Exploring Cancer's Fractured Genomic Landscape: Searching for Cancer Drivers and Vulnerabilities in Somatic Copy Number Alterations

Exploring cancer's fractured genomic landscape: Searching for cancer drivers and vulnerabilities in somatic copy number alterations

A dissertation presented
by

Travis Ian Zack

to

The committee on Higher Degrees in Biophysics

in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in the subject of

Biophysics

Harvard University
Cambridge, Massachusetts

July 2014

Thesis Advisor: Rameen Beroukhim Travis Zack

Exploring cancer's fractured genomic landscape: Searching for cancer drivers and vulnerabilities in somatic copy number alterations

Abstract

Somatic copy number alterations (SCNAs) are a class of alterations that lead to deviations from diploidy in developing and established tumors. A feature that distinguishes SCNAs from other alterations is their genomic footprint. The large genomic footprint of SCNAs in a typical cancer's genome presents both a challenge and an opportunity to find targetable vulnerabilities in cancer. Because a single event affects many genes, it is often challenging to identify the tumorigenic targets of SCNAs. Conversely, events that affect multiple genes may provide specific vulnerabilities through "bystander" genes, in addition to vulnerabilities directly associated with the targets.

We approached the goal of understanding how the structure of SCNAs may lead to dependency in two ways. To improve our understanding of how SCNAs promote tumor progression we analyzed the SCNAs in 4934 primary tumors in 11 common cancers collected by the Cancer Genome Atlas (TCGA). The scale of this dataset provided insights into the structure and patterns of SCNA, including purity and ploidy rates across disease, mechanistic forces shaping patterns of SCNA, regions undergoing significantly recurrent SCNAs, and correlations between SCNAs in regions implicated in cancer formation.

In a complementary approach, we integrating SCNA data and pooled RNAi screening data involving 11,000 genes across 86 cell lines to find non-driver genes whose partial loss led to increased sensitivity to RNAi suppression. We identified a new set of cancer specific vulnerabilities predicted by loss of non-driver genes, with the most significant gene being PSMC2, an obligate member of the 26S proteasome. Biochemically, we found that PSMC2 is in excess of cellular requirement in diploid cells, but becomes the stoichiometric limiting factor in proteasome formation after partial loss of this gene.

In summary, my work improved our understanding of the structure and patterns of SCNA, both informing how cancers develop and predicting novel cancer vulnerabilities. Our characterization of the SCNAs present across 5000 tumors uncovered novel structure in SCNAs and significant regions likely to contain driver genes. Through integrating SCNA data with the results of a functional genetic screen, we also uncovered a new set of vulnerabilities caused by unintended loss of non-driver genes.

Table of Contents

Front matter

There are a number of people who I would like to thank for their commitment to supporting me during the work discussed below.

First my wife, Jenn. I do not think I could have ever gotten through this if she was different in any way. Besides providing emotional support every day, Jenn has helped me with everything from paper editing and tissue culture when work was running late, to discussing my projects and being my most vocal critic (in the best possible way). As a scientist herself, she understood how projects can be quite consuming and never faulted me for late nights or an inability to stop thinking about science after I left the lab.

I'd like to thank my family for always supporting me in whatever I chose to do, especially my parents and grandparents, for believing in the importance of education. This includes my wife's family, who welcomed me without hesitation and supported my career decisions, even as I moved Jenn three thousand miles away and continue my attempt to go through life without a real job. In terms of support, I'd like to thank the Ashford family for financial support and a big final thanks goes to Jim Hogle and Michelle Jakoulov, for running, (really creating) a truly unique program. The flexibility and support is really beyond parallel and more than a student could ever ask for. I especially thank them for being understanding with my shifting career interest and my tardy, and somewhat backwards, approach towards degree completion.

On the science front, primary acknowledgement of course goes to Rameen, who has been a better mentor than I could have dreamed existed. From the start, he has let me jump into projects I had no business being a part of, and had the patience to give

me the time to learn so many different aspects of the science and techniques behind cancer discovery. I have learned so much from him about how to think productively about a scientific question, how to search for the best approach towards a solution, and how to efficiently tackle that solution. He also has been so patient with me. My communication skills were very under-developed (both in presentations and writing) when I entered his lab but he really took his mission as an educator seriously and gave me every opportunity to improve. Beyond his mentorship, he is a model scientist, both in his outlook towards the scientific pursuit and his ingenuity. Finally, his generosity is overwhelming, always making time to talk whenever I had questions or needed help and his passion for our work I think helped keep my energy for our projects going.

I want to acknowledge two hands on mentors in particular for their willingness to teach me from the ground up. Deepak Nijhawan and I worked in together on so many aspects of our project. For almost all aspects of experimental biology, working on our project was my first introduction to the techniques, even for things as basic as PCR. This usually meant experimental results were accomplished in parallel to traversing a learning curve, but I always felt Deepak was happy to be patient and teach me skills often only learned through trial and error. As our conversations often helped drive the experimental biology, he was also very excited about learning bioinformatics and discussing better ways to ask our questions. I do not think our collaboration would have been nearly as successful in a traditional compartmentalization, and I know I would have a much narrower skill set with which to answer future biological questions.

Similarly, Steve Schumacher has been my constant source of knowledge about bioinformatics and problem solving in software development. I knew a very small amount of programming when I joined the lab, but from day one, Steve has been my crutch and I knew I could always turn to him if I was stuck on something for more than five minutes. Besides just material help, he was always willing to be a sounding board with a critical ear, taking the time to listen to all my crazy ideas and thoughts, even if most of them were presented in a stream-of-consciousness fashion. In addition, his tireless dedication to seeing projects through to a conclusion, combined with his attention to detail, was a perfect complement to my ADD and disorganization.

Penultimately, though it is said a lot, it cannot be said enough: The unique and incredible atmosphere of support throughout the HMS and Broad community not just attracts great scientist, it breeds them. The ability to walk down the hall and annoy THE world authority in a given field or technique is so underappreciated as a formative process, and I think a culture that encourages this kind of openness starts at the top. When people see legends like Dr. Altsuhler, Dr. Lander, or the people on this committee willing to take time to teach and listen, they realize how important sharing skills and knowledge can be, and that maybe they have more time for it than they otherwise would have acknowledged. Personally, I would like to thank Drs. Meyerson, Getz, Hahn, Carter, Goldberg, Besche, Cherniack, Zhang, Tsherniack, Tamayo, Weir, and Rosenbluh and countless others for taking the time to help, teach, or listen to me, whenever I asked. Finally, a thanks to my committee for volunteering their time.

Motivation

Over 100 years ago Hansemann, and subsequently Boveri, described abnormal numbers and shapes of chromosomes as a distinguishing feature of cells derived from cancerous neoplasms[1,2]. As our understanding of chromosomes and their role in genetics has grown, this observation on the fundamental relationship between DNA and cancer has only been strengthened. It was not until 1960, that the first specific genetic defect associated with cancer was discovered, a chromosomal translocation in chronic myelogenous leukemia (CML)[3]. In the next couple decades, several genes responsible for driving tumor formation were slowly elucidated[4-7]. As genetic techniques became more refined, especially with the completion of the human genome project, the number of documented cancer genes and alterations has exploded[8-10].

The idea of cancer as an evolutionary process originating from a single cell is well established and has changed little in 30 years[11,12]. As in other evolutionary processes within finite populations, cancerous (or precancerous) cells acquire random genetic or epigenetic alterations, which may have beneficial, detrimental, or neutral effects on cellular fitness. Alterations that are strongly beneficial to cell survival and proliferation (driver alterations) may lead to expansion of the corresponding cell throughout the population (selective sweep). However, neutral or detrimental alterations present in the cell prior to these selective sweeps will also be present in the tumor (passenger alterations). Through this process, the genetic alterations observed at the time of diagnosis and/or tumor resection consist of some combination of driver and passenger alterations.

Early therapies postulated that known cytotoxic agents, such as mustard gas, would have similar effects on the unrestrained proliferation of cancer cells[8-11,13,14]. While cytotoxic agents are still frequently applied in cancer treatment, recent advances in genetics have introduced the promise of “targeted therapies”: treatment designed to target vulnerabilities caused by specific alterations present in cancer cells, but not normal tissue. While the basic concept of exploiting phenotypic differences between cancer and normal cells directly related to a genotype is a very general aim, in practice these efforts have focused on driver alterations directly involved in tumorigenesis. Because driver alterations are directly responsible for facilitating malignancy progression, it is reasonable to assume many may be required for tumor maintenance, and thus would represent a cancer-specific dependency.

Since the elucidation of the human genome, much of the research in cancer genetics can be broken down into three distinct pursuits:

1. Discovering driver alterations: Because we expect driver alterations to be preferentially maintained during the course of evolution, this often entails looking for alterations that recur in primary tumors more often than we would expect given our understanding of the background rate of the alteration.
2. Understanding how a driver alteration promotes or maintains cancer development: This step often involves experimental techniques to elucidate exactly how a genetic alteration alters cellular biology in a way that promotes tumorigenesis or tumor maintenance. Benchwork often represents both the backbone and a bottleneck of cancer discovery and has provided numerous insights not just into tumor formation, but basic cell biology and human physiology[15,16].

3. Therapeutic development: Once the biology of a specific driver alteration is understood, therapies (often small molecule inhibitors) are developed to exploit predicted dependencies based on this alteration. While the initial phase of this step is often carried out at academic institutions, much of the time and cost of this step involves molecular optimization and toxicity profiling, often at pharmaceutical companies.

Each patient's cancer likely has a small handful of alterations driving tumor development[17,18], meaning each patient presents with a very limited set of opportunities for driver-directed targeted therapy. While research in the causes and progression of cancer has led to innumerable advances in the understanding of cellular biology writ large[19-22], targeted therapies based on certain classes of driver alterations (such as tumor suppressors or transcription factors) have not always been easy to design or to implement successfully[23-25]. This leads to many patients in which we can identify the events that caused tumor formation, but cannot yet leverage this genomic information toward effective treatment. However, cancer dependencies predicted by driver alterations may only be a small subset of potential therapeutic targets. For example, driver alterations likely represent a small portion of all genetic differences between cancer and normal cells. By widening the search for cancer specific vulnerabilities beyond driver alterations, we may be able to open new therapeutic doors in targeted therapeutics.

One of the many forms of genetic alterations in cancer cells is changes in DNA copy number, which are collectively termed Somatic Copy Number Alterations (SCNAs). While typical mammalian somatic tissue is diploid (two copies of each gene), many

cancers exhibit wide deviations from this norm, ranging from complete loss (zero copies) of some regions to amplifying other regions hundreds of times. SCNAs affect a larger fraction of the genome in cancers than do any other type of somatic genetic alteration[26-28]. SCNAs can be caused by numerous distinct mechanisms, including multiple classes of errors in break repair and mitotic segregation[29-32]. SCNAs play critical roles in activating oncogenes and inactivating tumor suppressors[28,33-37] and understanding the biological and phenotypic effects of SCNAs has led to substantial advances in cancer diagnostics and therapeutics[38-41].

The following research focuses on uncovering contributions of SCNAs to cancer development and novel cancer dependencies. We have employed two complementary techniques. In the first, we used a large collection of primary tumors to characterize the patterns of SCNA across cancer and implicate specific regions likely to contain driver genes. In the second, we used an in vitro screen of gene dependencies to find genes that directly lead to cancer vulnerability when lost through SCNAs, regardless of whether they are driver or passenger alterations.

Chapter 1: Introduction

1. The analysis of copy number change

Introduction: Patterns of genetic alterations as a window into cancer biology

While a connection between certain behaviors and environmental conditions had been understood for centuries[42,43], it took until 1937 for scientist to prove a specific environmental toxin to be carcinogenic, with the exposure of mice to benzanthracenes, an element of coal tar. Concurrently, evidence arose of X-rays causing cancer[44] in addition to the (re-)discovery of cancer causing-viruses in the 1970s[45,46]. Combined with growing evidence of causative nature of these agents in genetic alterations[47,48] and ingenious work to show that cancers were derived from clonal expansion of a single somatic cell[49], it became evident that cancer was a disease of somatic, genetic alterations.

Therefore, a primary challenge in cancer biology is to distinguish the genetic alterations that allow clonal expansion and contribute to oncogenesis and cancer progression from the passenger events that are acquired during cancer evolution but do not contribute towards it[50-52]. The basic assumptions behind this pursuit are two-fold. The first element is that neoplasia is the result of a single-cell, evolutionary process that undergoes one or more clonal expansions. In such processes, genetic alterations that greatly improve fitness will be selected for and be enriched in subsequent measurements of the population (tumor biopsies, for example). The second assumption is that cancer biology remains relatively consistent across patients, meaning the same alteration may be selected for independently across multiple patients within a

population, and therefore would recur more frequently across biological samples than alterations with no selective advantage. These basic assumptions have, for the most part, proven fundamentally sound and have led to many breakthroughs in cancer, even before the introduction of high-throughput genetic techniques.

Without modern genetic techniques, early studies relied on qualitative visual analysis of chromosomal number and shape in neoplastic cells arrested in metaphase with colchicine[53,54]. While crude by today's standards, observation of static mitotic divisions allowed quantification and characterization of the chromosomal content, including "marker chromosomes" that were only observed neoplastic cells[55-57]. The principle finding in these early studies is summarized succinctly by Wakonig-Vaartaja[58]

"(a) neoplasia begins in a few or, more likely, in one cell, at least in certain types of neoplasia; (b) autonomous tendencies can be inherited from a cell to its descendants; (c) leukaemia, even with almost simultaneous symptoms in several tissues may originate at one site; from this it may spread through metastasis; (d) the majority of the cells in certain neoplasms are changed irreversibly, and not temporarily due to stimulation from other cells (physiological theories), otherwise the proportions of marked cells belonging to one clone would have been smaller; (e) the varying incidence of aneuploidy in the different sites, in spite of the same clonal origin of their cells, demonstrates genetic plasticity and random somatic evolution in neoplasms"

While less common, some "marker" chromosomes were observed repeatedly in patients of a particular disease, suggesting a role for these specific events in disease

formation[56,59]. However, identification of recurrent genetic alterations was limited by an inability to reliably distinguish between different chromosomes.

By 1970, experiments began to use quinacrine as a label for DNA regions for guanine-rich regions[60-62] and Giemsa staining for thymine-rich regions[63-66]. This, and other advances in cytogenetics, precipitated a series of conferences on cytogenetic nomenclature, culminating in the Paris conference of 1971, which set in place the basic nomenclature we use today[67], and led to a rapid increase in the discovery of DNA elements correlated with disease.

Besides more accurately describing previously discovered malignancy-associated alterations[3,4,68,69], the ability to accurately distinguish chromosome regions led to associations between specific regions and malignancy in a wide variety of disease states, even if the functional cause was as yet undetermined[70-74].

In parallel, the ability to laboriously sequence DNA[75] and amino acids[76], led to the discovery of many disease associated genes, with their genomic location often identified subsequently[77-80]. By studying tumor-associated viruses, researchers were able to isolate endogenous protein associated with transformation through biochemical isolation[81-83] or isolation of human homologues of viral proteins through transformation assays[7,83-86]. However, these discoveries were tethered to the relative simplicity of the viral genome, and the identification of human homologues, as well as their locations, was very tedious. Regardless, these first studies into the functional elements that lead to malignancy are responsible for discovery of many of the most well recognized tumor-inducing agents, including tumor suppressor genes(*RB1, TP53*)[22,87,88], oncogenes(*Myc, EGFR)*[89,90], viral components (hepatitis B, human papillomavirus)[91,92] and bacteria (*H.*

pylori)[93]. Again, in virtually all these cases, driver gene discovery was relatively opportunistic, discovering genetic elements through biochemical isolation or qualitative observation of highly recurrent genetic alteration.

A complementary approach, originally termed “reverse genetics” or positional cloning, attempted to map the gene responsible for a disease trait to its genetic location prior to evaluation of gene product biochemistry or physiologic function[94,95]. While traditional approaches strove to first isolate and characterize proteins associated with a phenotype, subsequently identifying its amino-acid and DNA sequence, these approaches involved challenging isolation and biochemistry techniques and often required model systems that adequately recapitulated the phenotype of interest. As an alternative, early genomics pioneers attempted to associate DNA regions with inheritance of a specific disease phenotype, slowly narrowing the phenotype-associated region down to a single genetic element either through comparison of the affected region in diseased versus normal cells[96-99] or through amplification and segmentation of the region through the use of a library of recombinant DNA vectors[100-102]. In this approach, only after a correlated genetic element is identified are experimental studies done to determine a mechanism behind its causal role. While also successful in mapping many human disease traits, the tedious nature of this approach precipitated the push for a complete mapping of the human genome[94,103], as well as new techniques for precise characterization of genetic differences between samples. One of the most influential characterization techniques was the DNA microarray.

History of DNA microarray analysis

The next big step towards large scale cancer genome characterization was the DNA microarray[104,105]. A microarray is the ordered, precision placement of many microscopic elements that allows segregation of signal or information in a very small system. For DNA microarrays, the segregated elements are DNA fragments ("target set"), covalently bonded to a solid support and then used to measure the amount of complementary DNA sequence that is fluorescently labeled in a small volume fluid sample ("probe set"). The advantage of this approach for interrogation of biological samples is multifaceted. A microscopic platform requires much less source material, while the high specificity of DNA to its complimentary sequence provides high signal to noise in a complicated sample, eliminating the need for step-wise sample simplification.

The first described DNA microarrays were developed in the early 1990s[105-107], and can be separated into 3 approaches, based on the construction of the "target" library.

In an early paper, cDNA was created from an *Arabidopsis sp.*, and a small set of putative genes were independently PCR amplified[105]. Small volumes of these PCR products were spotted in an ordered fashion onto a poly-L-lysine coated microscope slide[107]. After a chemical and heat treatment, the DNA probes were bonded to the slide surface and probed with fluorescently labeled mRNA samples. More generally, this approach used cDNA to construct Expressed Sequence Tag (EST) target libraries. While it is easy to interpret the direct readout from these cDNA libraries, they suffered from tedious library construction and selection bias against low abundance mRNA. Regardless, this approach was successfully used to measure gene expression in many systems[105,108,109]. In cancer, the ability to monitor changes in gene expression across a

large portion of the genome allowed diagnosis and classification of disease state based on broad patterns of expression[110,111].

A separate approach created target libraries by spotting cosmid or Bacterial Artificial Chromosomes (BAC) libraries onto a microarray[112]. They improved on some spatial resolution and quantification issues of previous approaches and were able to quantitatively assess levels of human DNA spiked into a full human genome, as well as high resolution comparison of chromosome 20 in normal and breast tumor lines[112]. These libraries were also laborious to curate and maintain.

Photolithographic synthesis of unique polymers in parallel directly on a solid state support[113,114] solves these problems of library creation and maintenance, by "printing" an array of short synthetic nucleotides of any conceivable sequence without need for an initial biological sample isolations step. Often termed "in-situ" DNA microarrays, this principle was demonstrated early in DNA microarray development[115], but the precise control over short target sequence made these arrays much better suited for sequence-identity questions[116-118] or discovery of disease associated single nucleotide variation[119,120]. Nucleic acid quantification was demonstrated with synthetic arrays in 1996[121].

While the utility of microarrays became obvious, the technical aspects and start-up cost of individual investigator-based microarray creation severely limited the wide-spread use of the technology initially. Because it was still impossible to characterize a whole genome on a single array, commercialization of microarray system led to individual development of target libraries, and microarray designs, which could be combined to answer specific biological questions (reviewed in Bowtell, 1999)[106]

Microarrays and copy number alterations.

Though now nearly universally performed with microarrays, comparative genome hybridization (CGH) was originally developed as a cytogenetic technique[122] in parallel to microarrays. Tumor DNA labeled with biotin was combined with normal genome reference DNA labeled with digoxigenin in equimolar amounts. This solution was then hybridized to metaphase spreads of normal cells, and the fluorescent signal intensities derived from each sample (tumor vs. normal) at each locus along the chromosome was used to determine relative copy number at these loci[123]. As signal was derived directly from condensed chromatin, resolution was limited[124] and gene identification again required subsequent, more narrow genetic interrogation[125,126]. In addition, accessibility and hybridization efficiency vary widely across the genome during metaphase leading to interpretation difficulties. But the idea behind CGH transferred very well to discovering copy number abnormalities using microarrays. Initial array-based CGH[112,127] were based on BAC (or similarly constructed) target libraries and improved the resolution of traditional CGH techniques enormously. Subsequently, EST libraries were used to probe cancer copy number changes[104] but in both cases, copy number analyses had difficulty replicating the success of EST-based mRNA expression studies, as copy number changes are typically more subtle, and thus harder to detect, than the changes frequent in expression patterns, which can often span orders of magnitude.

The application of synthetic oligonucleotide technology to array CGH would only come after the human genome project[128,129]. Agilent Technologies used the consensus sequence to create an array of over 20,000 60-mer oligonucleotides that tiled 4

chromosomes[130]. While this platform still had a low signal to noise ratio, the high density of probes in each region allowed for signal intensity across adjacent probes to be averaged, improving resolution to the 50-100kb range and leading to more accurate interpretation of copy number changes.

While homozygous loss is a common method of tumor suppressor inactivation, loss of heterozygosity (LOH) can also inactivate tumor suppressors[131,132] if the remaining allele is rendered non-functional by some other means. While early aCGH techniques may not have been sensitive enough to accurately distinguish between one and two DNA copies, Mei et al[133] and Lindblad-Toh et al[134] both realized one could take advantage of single heterozygous single nucleotide polymorphisms (SNPs) to increase the information from each oligonucleotide and more easily locate regions of the genome that had suffered LOH[134,135]. These methods improved as more germline SNPs were identified and added to microarray detection[136,137], eventually allowing for quantitative models of copy number change detection[138,139].

Microarray information processing

While powerful, microarrays are subject to a number of different sources of error, and the algorithms used to determine somatic copy number changes from microarray platforms have continued to improve. Systematic noise in microarray experiments is primary derived from two sources, the first relating to the target library and the second relating to the probe set[140]. DNA hybridization relies on the high specificity of complementary DNA sequences to bind one another, even in a complex solution. Each unique sequence represents a different chemical reaction, with different reaction rates,

specificities, and equilibria. As such, signal comparison across different targets on the microarray is very challenging. From the perspective of sample processing, most DNA samples require PCR amplification prior to genomic analysis. PCR amplification comes with the same set of chemical biases between reactions as the target sequences. However, because amplification is done over multiple cycles in series, errors and sequence biases are propagated and amplified with each reaction, leading to huge artifacts in analyses if not properly accounted for[140]. While gene expression arrays experience these sources of noise as well, the basic goal of gene expression analysis is to quantify the same molecule (mRNA species) across different samples. Differences in chemistry across different molecules are therefore only of secondary concern. In contrast, in DNA copy analysis, we wish to use signals from multiple target sequences, all derived from proximal areas of the genome, to inform the total copy level at that locus. This requires comparison of different chemical reactions, and therefore requires more sophisticated analyses.

For a single target sequence, we would expect a linear relationship between the amount of complementary DNA in a sample and intensity of signal from that location on the microarray. This is indeed the case for moderate levels of copy level change (0-5) [138,139], while fluorescence saturation is observed for very high levels of copy number amplification (>100)[138].

In the initial analysis of tumors, correlation between adjacent genomic markers within a given sample was low. In Bignell, 2004[138], rather than providing relative genome-wide copy level, copy level determination was restricted to either small areas with multiple and concordant, unlikely intensities, or to very large areas with unlikely

average intensities. They found 24 regions with at least three consecutive SNPs with outlier intensity values when compared to a diploid sample, suggesting high level copy number change (putative regions of high level amplification or homozygous deletion). For more subtle copy number changes, averages of fluorescent intensity over large regions were able to predict whole chromosome changes to a limited degree.

Around a similar time, Zhao, 2004[139] developed more sophisticated data analyses of copy number change providing better opportunities to recognize CNA. They modeled the signal derived from the array as a Hidden Markov Model (HMM), with each target sequence informing the probability of a real copy level change in neighboring targets. In brief, the proportion of the genome at a given copy level is initially based on heuristics (~90% diploid, with more extreme copy levels progressively less likely). The HMM then tried to determine where copy level changes occur (and hence the copy level at each locus), with transition probabilities based on both these proportions and the relative genomic spacing of target sequences. Finally, this process is repeated with new estimates for the proportion of the genome at each copy level to refine transition and copy level probabilities. Using this more quantitative assessment of copy level changes, they were able to find smaller regions of subtle copy level change in many cell lines, as well as high level gains in many known oncogenes.

These initial efforts at signal detection recognized that even if large portions of the genome are affected by copy number change, transitions between different copy levels are relatively rare. This means that signal from adjacent sequences reinforce each other, and utilizing this increases signal to noise. Going further, a systematic method for noise averaging that leads to an estimated copy number value at each

locus, as well as the most likely genomic position that signifies the start of copy number change (often referred to as copy number breakpoints) serves as a foundation for finding significant recurrence of copy number alterations. Circular binary segmentation (CBS) does this by taking each segment (starting with the whole chromosome) and recursively asking if there is likely to be a copy number breakpoint within that segment[141,142]. After finding the most likely breakpoints, it then averages intensity values across each segment to create segment-level data.

To improve resolution, commercial arrays were developed to contain both common SNPs and additional synthetic oligonucleotides tiled across the genome (copy number probes) [143,144]. The sequence of complementary sample DNA is (presumably) homogenous for these target sequences, therefore leading to different signal models. As projects moved to these platforms, decreasing cost and increasing reproducibility, algorithms were created and refined to integrate information from both SNP and copy number targets[145]

Analyzing DNA derived from sampling a continuously evolving, heterogeneous population comes with its own complications. As primary solid tumor samples are traditionally derived from biopsies or resected tumors, DNA derived from all cells in these samples not only can mask tumor heterogeneity, because samples invariably have some mix of tumor and somatic tissue, normal cells can dampen tumor signal. Interpretation can also be complicated by large deviations from diploidy. Because cell number is not assessed prior to sample processing, a perfectly tetraploid cell would look identical to a diploid cell from a microarray's perspective. These two confounders (tumor purity and ploidy) limited copy number determination to assessing relative changes

across loci, and complicated comparison across samples. Recently, several algorithms[146,147] have attempted to use concordance in copy level across segments to infer tumor purity and ploidy and obtain the clonal number of DNA copies of each locus. ASCAT[146] was the first method introduced for predicting purity and ploidy. It does so by looking for the best solution that explains the total copy ratio and relative ratio of the A/B allele outputs at each locus, using purity and allelic ploidy as parameters. Then, to find a solution for purity that minimizes these local solutions globally, they find the purity value that minimizes the overall distance between the copy number of each segment and its nearest, non-negative integer. The ABSOLUTE method used a similar underlying model, with improvements by allowing the flexibility of taking in other information, such as subclonality, somatic point mutations, and karyotypic probability based on previously characterized tumors[147].

SCNA structure in cancer

Cancer genomes arise through a diverse set of permanent (or semi-permanent) alterations that can be separated into broad categories: sequence alterations, balanced structural rearrangements, copy number alterations, and epigenetic modifications. Sequence alterations are somatic changes in DNA sequence and include single base substitutions as well as small insertions and deletions that affect a few nucleotides. In contrast, epigenetic modifications, which encompass DNA methylation and histone modifications, are changes relative to the natural state of the cell of origin. As the precise cell of origin, much less its epigenetic state, can be difficult to determine, epigenetic modifications can be difficult to interpret[148]. Finally, structural rearrangements

and copy number alterations are reorganizations of the DNA content in a given cell, which can lead to changes in sequence (as disparate DNA elements are joined together) as well as changes total number of copies of affected regions[33,149].

Mitosis associated copy number alteration

Whole chromosome instability (W-CIN) is generally thought to arise from either errors in the centrosome duplication cycle, leading to multiple centrosomes during mitosis[150], or errors in spindle assembly checkpoint, which can allow progression of anaphase even without a full complement of kineticore attachment[151]. The centrosome is the organelle responsible for organizing and anchoring chromosomes involved in mitotic DNA segregation, and as such, its replication is intimately connected to the cell cycle and start of DNA replication[152]. Cell cycle components, particularly *CDK2* and *Rb*, are important in the initiation of centrosome duplication and are also frequently dysregulated during cancer development. Errors in centrosome duplication can be a byproduct of driver gene activation or inactivation of these cell cycle components[153-155]. Over-duplication of centrosomes may also occur after abortion of normal cell division, followed by aberrant re-entry into S-phase through inactivation of the p53-mediated checkpoint pathway[156]. Unlike the first mechanism, this route will lead to coinciding duplication in DNA for the cell and its progeny (tetraploidy). While there are mechanisms in the cell to prevent mis-segregation when microtubules from multiple centrosomes bind a single kineticore[157], these mechanisms get overwhelmed in the case of over-duplication of centrosomes and can lead to subsequent aneuploidy[158].

Whole chromosome instability is also frequently the result of errors in the spindle assembly checkpoint (SAC). While this checkpoint is active, a single unattached kineticore is enough to prevent initiation of anaphase[159]. Perhaps because loss of many of the genes in the complex directly responsible for this delay[160] is embryonic lethal[161-163], they are infrequently mutated in cancer[164]. However, dysregulation of this checkpoint through other means is fairly common, such as alterations in APC[165] which can stabilize microtubules and REST[166] and VHL[167], which regulate components of the checkpoint complex. More recently, recurrent alterations have been observed in the Protein phosphatase 2 complex[168], which regulates the attachment of the anaphase-promoting complex to the mitotic spindle[169].

In addition to chromosomal mis-segregation, which has been well documented from the beginning of cancer cell biology[170], breakage-fusion-bridge cycles (BFB) and chromothripsis represent processes that result in sub-chromosomal alterations that are believed to be associated with anaphase in mitosis[30,171]. BFB cycles are thought to originate from an initial dsDNA break during anaphase, leading to loss of the telomeric region[172]. After subsequent replication, the corresponding sister chromatids both lack telomeres, leading to fusion of the two strands and a “bridge” between them during anaphase. This bridge is only resolved when the centromeres of the sister chromatids are pulled to opposite ends of the dividing cell, leading to a mechanical break and again revealing chromosome ends uncovered by telomeres[173]. This leads to a series of telomere bounded amplifications or deletions. Chromothripsis is a recently described phenomenon where a single chromosome undergoes dozens of genomic rearrangements, including the inversion or deletion of many regions of DNA, in a single

complicated event[174,175]. Though the mechanism of chromothripsis remains unclear, it has been postulated to occur while the chromosome is condensed during mitosis, either through temporary chromosomal isolation, mechanical stress during anaphase, or external mutagens, such as ionizing radiation[30].

Mechanisms of copy number alterations during DNA replication and break repair

Common features of cancer cells include DNA damage and dysregulation of cell cycle processes[17]. As replicative stress can cause frequent dsDNA breaks during progressing replication forks, both of these processes lead to increased requirement for DNA break repair. Errors in the repair of these dsDNA breaks can lead to local changes in DNA copy number, through either homologous or non-homologous recombination[176]. Homologous Repair (HR) is the less error prone of the two repair pathways and requires long tracts of sequence homology near the breakpoint to be repaired[177]. The repair process involves invasion of the homologous sequence with a single strand of the broken DNA, and uses it as a template to extend beyond the break. Conversely, non-homologous repair is relied upon when extensive sequence homology cannot be found as a template. This process relies on micro-homology, if available, or can continue with no homologous region at all and, theoretically, is much more likely to lead to copy number changes.

As its name suggests, homologous recombination requires a template strand with significant local homology to the double stranded break. If both sides of the DSB are present (3' and 5'), repair can pass through either a double Holliday junction intermediate, or through synthesis dependent strand annealing (SDSA). In the former,

template strand invasion is followed by resolution of a double Holliday junction, which can lead to allelic crossing over[176]. Even if alleles are correctly paired during HR through Holliday junctions, crossing over events can lead to loss of heterozygosity (LOH) if chromatids carrying the same allele segregate together during mitosis. Alternatively, while SDSA also involves template strand invasion, the ssDNA molecule from the DSB strand is separated from the template and re-ligated to a newly uncovered 3' overhang from its original DSB partner, preventing the possibility of a crossing over event[178]. If HR incorrectly pairs two regions of homology (not directly opposed alleles), crossing over can lead to inversion of duplication, depending on the orientation of the homologous regions.

If a DSB occurs during replicative fork propagation, resulting in only a single DSB, repair proceeds through the Break-Induced Replication (BIR) pathway[178]. This pathway proceeds through a series of strand invasions, separations, and extensions until a high processivity replication is formed to continue DNA replication. While this process is immune to crossing over, LOH can occur if the DSB strand invades the homologous chromosome instead of the sister chromatid. If HR takes place across two non-allelic homologous regions, it can lead to translocations[179] and potentially other copy number changes.

Non-homologous end-joining (NHEJ) and microhomology mediated end joining (MMEJ) are the two main repair pathways that do not require long sequences of sister-strand homology. Without this strict template strand requirement, these can often lead to copy number alterations[180,181]. As its name implies NHEJ attempts to relegate two DSB; errors can often lead to micro-deletions or insertions. Alternatively, MMEJ tries to

find small regions of homology (5-25 nucleotides) near the DSB by resection to expose ssDNA until a region of microhomology between the two dsDNA elements are found, followed by ligation and synthesis. This process can lead to deletions of variable length, and potentially translocations if DSB are prevalent in the nucleus.

As many putative mechanisms for copy number alteration are known, there have been many efforts to identify mechanistic correlates of SCNA. One of the more important revelations from such studies has been the role genome architecture has in observed structural rearrangements and copy number changes[182]. Spatial organization during interphase seems to be related to frequent individual translocations[183,184], though whether this is due to increased likelihood of cross-over errors after DSB or due to similar times in DNA replication remains unclear[185,186].

Significance analysis in somatic copy number change

Having characterized many independent cancer genomes, an important goal is to discover recurrent alterations that lead to tumor formation, ideally taking into account what we know about the mechanism of their generation. The tools to do this are constantly evolving. In many of the first genetic abnormalities identified, an alteration is so ubiquitous across a particular disease, its significance can be determined qualitatively[3,187,188]. Unfortunately, in most solid tumor types, a ubiquitous alteration, pointing to a specific and unique cause for malignancy, is lacking. This, combined with the large portion of the genome altered in each sample, made a statistical approach towards the discovery of recurrent alterations a necessity to discover the vast majority of cancer genes.

In copy number analysis using DNA microarrays, initial methods for identification of significant regions were straightforward. One such approach involved finding the minimal common region of overlap between copy number changes in different samples, then refining the list of interesting regions through heuristics[189]. Unfortunately, the results of early methods designed to discover regions containing potential driver genes were very dissimilar, even in parallel datasets within the same disease[50,137,189]. To address this, multiple groups undertook projects designed to improve our ability to identify regions likely to contain driver alterations[50,190-192] (reviewed in [Rueda 2010][193]).

The main advantage of these techniques over the initial attempts was propositioning a background rate of SCNA and using this proposed distribution to find regions with more alterations than expected. For example, the approach developed by our laboratory, Genomic Identification of Significant Targets In Cancer (GISTIC) creates a score for each locus by averaging the copy level at each locus across samples, then determines the probability of obtaining that score assuming a random distribution of copy levels within each sample[50]. RAE[190] applied a similar method, with the addition of a noise model of each individual tumor sample and boundaries.

Building on its immediate predecessor[50], GISTIC 2 recognized that most loci affected by copy number alterations are from chromosome-level changes and specifically looked to separate the significance of chromosome-level SCNAs from shorter “focal” SCNAs[1,2,26]. Unfortunately, the frequency and extent of chromosomal alterations often obscures recurrence of focal copy number alterations. To resolve this issue, GISTIC 2 introduced an algorithm that attempts to separate chromosome level events (“broad” events) from focal events, and assess the regions of significant

recurrence for each of these events separately[3,194]. Assigning significance between broad and focal events independently was a great leap forward in the discovery of driver SCNAs, leading to its broad implementation across the field.

The biological complexity of primary tumor samples can complicate copy number analyses using DNA arrays. DNA microarrays measure the amount of DNA at a given locus, relative to the total amount of DNA extracted from the sample, providing information about “relative” copy number changes. For example, the microarray signal from a clonal population of diploid cells would be identical to one of tetraploid cells. This means there is potential for lost information, even in clonal populations. Moreover, the ideal case of an entire sample of clonal tumor cells is rare. Instead, tumor samples often contain large amounts of normal somatic tissue, such as infiltrating inflammatory cells or cells from the host organ. In addition, as a continuing evolutionary process, the cancer cell population is inherently heterogeneous. All these genetic backgrounds are homogenized and placed on the array simultaneously, which obscures the signal of clonal events. Carter et al.[147] designed a series of algorithms to address these issues by modeling the most likely values of tumor purity, ploidy and heterogeneity to lead to the observed data.

Using genetic patterns to imply functionality

Even once a region is determined to be significant for SCNA recurrence, the challenge of elucidating the functional unit responsible for this significance is often only partially complete. Unlike many other somatic alterations, the genetic footprints of SCNAs vary widely, from small regions of a single gene, up through whole chromosomes (100s to 1000s of genes). This means that even if we identify a specific

region as being recurrently affected, the genetic target of these alterations may remain ambiguous without additional information. There are many potential approaches towards removing this ambiguity.

In traditional approaches, functional consequences of copy number change are assessed for each element within the region through molecular biology, such as *in vitro* or *in vivo* ORF and shRNA transformation or viability assays. While experimental approaches are, in some ways, the gold standard, the process can be very low throughput, with each gene having to be independently tested. In addition, each experimental system comes with its own caveats. Assays are designed to test for specific tumor-promoting properties, such as transformation[195] or invasion[196]. Using the incorrect assay, could therefore lead to a false negative result for driver genes with an untested tumorigenic function. Even if the correct test is run, many tumor-promoting effects require a specific genetic context to lead to tumor promotion. These contexts can be difficult to predict and/or replicate in a laboratory setting[197]. Finally, some driver alterations, such as TP53 or BRCA1, will not transform cells in isolation because they disable growth checkpoints or increase the likelihood of other driver alterations, working indirectly to promote tumorigenesis[198,199].

For these reasons, using the genetic context of a driver alteration to both limit the number of hypothesis and refine the functional question can be very useful. For example, positive correlations with other genetic events may indicate functional synergies, where one is beneficial only in the context of the other. Alternatively, anticorrelations may indicate functional redundancies as redundant events would not be

required in the same primary tumor. Several approaches have been developed to determine functional effects of genetic events based on anticorrelation patterns.

In the past, using correlative structure in the genetic alterations of cancer could be separated into two distinct goals. In the first, the goal is to use biological pathway information, combined with genetic correlations, to improve our power to detect cancer-driver genes. In the second, the goal is to leverage genetic correlations between known driver alterations to inform function.

Examples of the first approach are exemplified by work form the Rafael lab[200,201]. In this approach, rather than identify significantly mutated genes, investigators identify significantly mutated subnetworks and the genes involved in them. In this way, they choose not just genes that are frequently mutated, but a set of genes that are both frequently altered AND that tend to be more mutually exclusive than expected by chance (and hence potentially explain more tumorigenesis).

The second approach looks to inform functionality of known driver alterations by looking at their patterns of mutual exclusivity[202]. Standard tests of mutual exclusivity have proven inadequate for significance analysis in alteration datasets. Developing more sophisticated algorithms have led to interesting biological relationships[203].

However, these techniques have not been systematically tested and implemented in the setting of SCNAs, where the structure and patterns of alterations may provide unique challenges. While permutation analyses are a standard approach for data that does not adhere to canonical distributions, SCNAs, unlike other alterations, often span large and overlapping regions, making a permutation analyses that maintains event structure challenging.

Cancer discovery as a collaborative goal

Competition can be a great motivator of progress and discovery[204] and the turn of this century saw multiple, parallel initiatives designed to collect and analyze collections of primary tumors from the same disease. It was recognized that, while these were very successful independently, collectivizing efforts and datasets would lead to more power in pattern detection and, ultimately, disease understanding[205]. In 1997, Vice President Albert Gore announced the Cancer Genome Anatomy Project (CGAP) website[206,207], with the goal of making public 1. a large, standardized collection of genetic data derived from primary tumors and 2. state of the art technological tools to analyze this dataset. As the majority of this project took place prior to the completion of the human genome, early results focused on cytogenetics and gene expression profiling[205,208], with whole genome structural information integrated subsequently[209]. Subsequently, The Cancer Genome Atlas (TCGA) was a collaborative effort instituted by the NIH to provide a comprehensive genomic characterization of 20 of the most common cancers in the United States, as well as smaller characterization projects in many, less common diseases. This initiative has been widely successful and already produced landmark papers in a number of common diseases[5,6,210 9,10,211 11,12,168,212]. Within the TCGA, the Pan-Cancer project was a collaboration to characterize and compare 11 of the most common diseases with the aim to identify patterns across all cancer subtypes, as well as distinguishing features of each disease.

2. Non-driver dependencies as a means of cancer therapy

While cancers arise from the accumulation of somatic genetic only a subset of these alterations ("driver events") promote malignant transformation by activating oncogenes or inactivating tumor suppressor genes. Many somatic genetic alterations are the consequence of increased genomic instability that occurs in cancer but do not contribute to tumor development ("passenger events"). Even among events that promote tumor development, many of these alterations, specifically SCNAs, affect multiple genes simultaneously, with presumably a limited number of affected genes driving cancer ("driver genes") while the rest are simply tolerated by the now more genetically fit cell (bystander genes).

The demonstration that cancers are often dependent on specific driver oncogenes has stimulated efforts to find and exploit these targets therapeutically. By definition, driver events promoted cancer cell growth or development at some point during evolution, so expecting many of them to be specific cancer dependencies at the end stage of disease is a logical hypothesis that has proved successful therapeutically. For example, cancers that harbor translocations such as BCR-ABL or EML4-ALK, or mutations such as EGFR or BRAF, well proven transformative alterations, have been shown to depend on the activity of these gene products for tumor maintenance[213-215]. Indeed these alterations have been successfully targeted in the development of novel targeted inhibitors in a number of cancers including leukemia, lung cancers and melanoma[216-218]. Therefore, the presence of such an alteration often predicts response to drugs that inhibit the function of these proteins[38,217-219].

However, the focus on driver alterations as therapeutic targets presents some limitations. For one, many driver alterations belong to classes that have been very difficult to target with small molecule inhibitors, such as transcription factors or loss of function alterations in tumor suppressors[220,221]. As most primary tumors may be dependent on only a small handful of driver alterations[222-225], focusing solely on driver vulnerabilities leaves each patient with just a few opportunities for targeted therapy, even in the most ideal case.

An alternative strategy to target cancers is to target genes that are not oncogenes, but which cancers require to withstand cancer-specific stress[226,227]. In comparison to normal cells, cancer cells rely inordinately on pathways that abrogate a variety of cancer related stressors, including DNA damage replication stress, proteotoxic stress, mitotic stress, metabolic stress, and oxidative stress[228]. Even though proteins within these pathways may be essential in all cells, pharmacologic inhibition may create a therapeutic window as a result of a cancer-specific stresses.

While looking for pathways under increased cellular requirement in cancer is one approach at the discovery of non-driver dependencies, another is to look for dependencies caused by the large-scale rate of genetic alterations in cancer. For example, the majority of human cancers harbor copy-number alterations involving the loss or gain of broad chromosomal regions and these may be either driver or passenger events. As each chromosome arm can contain hundreds to thousands of functional elements, even driver chromosomal SCNAs affect a large number of bystander genes. In fact, the idea of chromosomal aberrations that drive cancer while affecting bystander

genes was recognized 100 years ago this year, in the original paper linking chromosomal aneuploidy to cancer[1]:

“Another possibility is that there is a specific inhibitory mechanism in every normal cell that only permits cell division to take place when this mechanism is overcome by some special stimulus. It would accord with our basic concept if one assumed that there were specific chromosomes that inhibited cell division. If their inhibitory effect were transitorily overcome, then cell division would resume. A tumor cell that proliferated without restraint would be generated if these `inhibitory chromosomes' were eliminated. In this case, the tumor cell would also lose all the attributes that were located exclusively in the same chromosome as the inhibitory factors.

However, the hypothesis that there are chromosomes that stimulate cell multiplication is also compatible with our proposal. In this view, cell division would take place when the operation of the stimulatory region of the chromatin, normally too weak, is enhanced by some active agent. The unrestrained proliferation of malignant tumour cells would then be due to a permanent excess of these stimulatory chromosomes.”

A dependency induced by the loss of such passenger and bystander genes in primary tumors was postulated 20 years ago as potential indirect vulnerability in cancer cells[229]. The hypothesis postulated that random passenger alterations would lead to decreased abundance of certain essential proteins, which would render cancer cells highly vulnerable to further suppression or inhibition of those genes.

In contrast to driver events, vulnerabilities associated with passenger or bystander alterations are much more difficult to detect through genetic enrichment

alone. As driver alterations promote tumorigenesis, they are likely to recur across multiple samples, and are more likely to predict dependency in samples in which they do occur. Incidental alterations, on the other hand, must be well tolerated, or at most, only slightly deleterious. However, in the absence of a positive selective pressure there would be very little evidence for dependency on bystander alterations in genetic data alone. An exception to this would be relative genomic integrity at a given locus, suggesting negative selection in cells harboring this alteration. Unfortunately, this would correspond to dependencies in a very small subset of patients (as the alteration is even rarer than the expected rate). Thus, ideally we are looking for passenger alterations that predict dependencies, yet do not lead to significant decreases in fitness at the time they are acquired.

3. shRNA screens to discover cancer dependencies on non-driver SCNAs

Expanding cancer genomics with functional screens

While genetic analysis of primary tumors can elucidate a lot about a cancer's evolutionary history, it has its limitations. For most datasets, primary tumors represent a window into the state of a tumor at one point in time. Although factors such as genetic context can provide clues as to the functional consequences of these alterations, these are best examined with direct experimentation.

High-throughput screens of cancer cell lines can directly ascertain the functional consequences of genetic or environmental perturbations. These can complement genetic analysis as a hypothesis-generating tool for discovery of cancer dependencies. The most common screens involve small molecule perturbations, phenotypic perturbations, or genetic perturbations[230-232]. One example, Project Achilles, was

developed to discover novel cancer dependencies by leveraging the full power of RNAi in a carefully constructed high-throughput system designed to assess relative gene dependency across a large panel of well-characterized cancer cell lines. In brief, this projected attempted to assess cell viability in 100 cell lines after suppression of over 11,000 individual proteins using a pooled library of shRNA lentiviral vectors (with each clone harboring an identifying barcode)[233]. To generate the shRNA library, a set of rules curated from the literature were implemented to improve knock-down efficiency and avoid obvious off-target effects[234]. Cell culture techniques and data processing are also outlined in Cheung et al[233].

The raw result of this large project is relative viability of cells infected with a specific shRNA, as compared to cells containing other shRNA in the lentiviral pool, as measured by relative abundance of a corresponding DNA barcode after a long period of cell growth. Given the efficacy and off-target concerns outlined below, interpretation of this data is challenging. In anticipation of these concerns, a level of redundancy was engineered in the screen, with each gene being targeted by at least three independent shRNA sequences, with a median number of five shRNAs per gene. This increases the chance that multiple shRNAs will effectively knock down the target allowing us to separate microRNA (miRNA) mediated off-target effects, which would likely be unique to each shRNA, from on-target responses, which would likely be shared across all effective shRNAs targeting a given gene [discussed below]. Separating these two effects has been accomplished a number of different ways and is a continually evolving problem.

shRNA: power and cautions in high-throughput screens

Over the past 10 years, the use of short interference RNA (siRNA) to suppress translation of specific proteins has become a powerful tool in molecular biology. Originally discovered as an external modulator of gene regulation in *Caenorhabditis elegans*[235], and subsequently extended to higher organisms[236,237], this system uses 21-23 nucleotides of double stranded RNA (siRNA) to target and eliminate longer strands of RNA (typically mRNA) containing the same base pair sequence[238]. In vitro RNAi can be utilized by either transfected double stranded RNA fragments, or transduced cells expressing an RNA hairpin which functions in the same manner (shRNA). In cells transduced with dsDNA, the mRNA product forms a looped secondary structure between complementary sequences that is cleaved by Dicer before being loaded onto the RNA-Induced Silencing Complex (RISC)[239]. Double stranded RNA is unwound into a passenger strand (which is degraded) and a guide strand, which is incorporated into RISC. RISC then uses the loaded guide strand to locate and target complementary mRNA for degradation, using its intrinsic Argonaute component[237,240].

The exogenous addition of siRNA/shRNA to suppress specific protein function is widely used, but its use is not without caveats. These can broadly be classified as shortcomings in either efficacy or precision. Both the siRNA sequence and the location of that sequence in the target mRNA can greatly affect the efficiency of translational suppression. Most mRNAs are composed of hundreds of base pairs, offering a large number of potential siRNA designs. However, as this technique has progressed, it has become increasingly clear that a significant percentage of siRNAs designed to suppress a given protein will fail to do so to any appreciable degree[241,242]. A set of “guidelines” have been discovered that improve the chances of a designed siRNA having a

suppressive effect, but these methods do not guarantee functionality[243-245]. With this in mind, validation of efficacy in each experimental system, by western blot or similar procedures, is a requirement for the use of siRNA. While this is straightforward during low-throughput experimentation, the variability in efficacy between different siRNA is a primary challenges in design and interpretation of hypothesis generating high-throughput RNAi screens.

In addition to issues with efficacy, siRNAs frequently generate many unintended cellular perturbations, or “off-target” effects[246]. The easiest off-target effect to understand, predict, and avoid is siRNA-mediated decay of mRNA sequentially similar to the target. This can be predicted based on BLAST alignment and avoided by careful selection of the target region, avoiding sequences that are shared with other functional elements of the genome. However, the siRNA pathway shares many elements with a similar, yet distinct, pathways of RNA-mediated translational regulation termed micro-RNA (miRNA)[247]. Rather than recognizing an exact sequence, the miRNA pathway recognizes imperfect homology between the 3’UTR of an mRNA and a “seed region” on the dsRNA[247]. This allows a single endogenous miRNA to regulate multiple mRNA containing slightly different sequences, which is great for biological efficiency, but complicates interpretation for many carefully designed experiments. The target and strength of these off-target effects can be very challenging to predict[246,248,249] or even determine experimentally[241]. In a low-throughput setting, a causative relationship between suppression of the target protein and an observed phenotype is verified using a series of well-established controls, including multiple target sequences, scrambled sequences, and rescue experiments. Unfortunately, as in efficacy issues, these simple

controls in a low throughput setting quickly become challenging in the setting of a high-throughput screen.

The high potential for false positives and low signal to noise means the design and execution of a high-throughput RNAi screen is only half the battle. Careful quality control, thoughtful data analysis, and experimental validation of individual results are just as important.

Generating high-confidence gene-dependency scores from shRNA pooled screening data across a large panel of cell lines.

Given a set of cell lines and data on relative viability for each cell line in response to numerous conditions, one standard hypothesis is that functional dependency can be predicted by a specific genetic alteration. In this case, by segregating cancer cell lines into two categories based on presence of the genetic alteration in question, you can look for functional perturbations, such as effect of gene suppression or drug tolerance, that best separate these two classes from one another.

Besides allowing us to start with a testable hypothesis, pooling samples we expect to exhibit the same phenotype renders our hypothesis less susceptible to experimental and biological variation, allowing us to spot trends in viability that may point to underlying biology. Because most “normal” somatic tissue does not grow *in vitro*, susceptibilities found in this assay use a large panel of genetically diverse cancer cell lines as controls and allow us to immediately discount dependencies shared by all cells.

Finding gene dependencies using shRNA in the two-class comparison setting has been accomplished in a number of different ways. Most importantly, each method

must account for both the efficacy and the off-target effects described above, and does so by integrating information from all shRNA designed to target a gene. The most basic method frequently used is the “second best” shRNA approach[250,251]. In this approach, we score genes by the second most effective shRNA in differentially separating the two groups. This accounts for the possibility that the effect of the strongest shRNA *may* be miRNA mediated, while also realizing that the number of ineffective shRNA (#s 3-X) may vary across genes. However, this incorporates very little information about the distribution of all shRNA for a given gene. RNAi Gene Enrichment Ranking (RIGER)[250,252] ranks each gene by enrichment of its shRNAs towards differential susceptibility between the altered and unaltered groups for each shRNA. While RIGER combines information from all shRNA in the screen, it systematically penalizes genes with multiple ineffective shRNA, even if they also have multiple effective shRNA providing good signal.

Shao et al.[253] developed an algorithm to analyze pooled shRNA data that improves on the shortcomings of the previous two approaches. The basics of this approach are illustrated outlined in [Shao et al.][253]. If the viability changes in response to two independent shRNA sequences are related to the same gene, we would expect these changes to track similarly across cell lines within our dataset. This idea of searching for similar patterns in shRNA targeting the same gene would theoretically ignore ineffective shRNA, while simultaneously removing shRNAs whose effects are dominated by the off-target miRNA pathway. This approach removes these shRNAs to create a “gene-dependency” score based on corroborating shRNA targeting a gene and removes genes that are not represented by multiple consistent shRNAs.

Chapter 2: Structure of somatic copy number alterations

Section Goals

Determining the patterns of somatic copy-number alterations (SCNAs) and how they promote cancer is important to understanding the disease. We characterized SCNA patterns among 4934 cancers from The Cancer Genome Atlas Pan-Cancer dataset and have integrated rigorous statistical approaches into these analyses, including absolute allelic copy-number profiling, as well as novel computational tools to determine individual SCNA events and their temporal ordering from these profiles. Whole-genome doubling, observed in 37% of cancers, was associated with higher rates of every other type of SCNA, with most of these events occurring subsequent to genome duplication. SCNAs that were internal to chromosomes tended to be shorter than telomere-bounded SCNAs, suggesting different mechanisms of generation. Finally, we developed a method for predicting chromothripsis using SNP6.0 arrays alone and found that specific chromosomes and diseases, including chromosome 9 and 12 in glioblastoma multiforme, were significantly enriched for these events. We present this dataset and analyses as a foundation for further analyses.

Methods

1. Event Decomposition

We deconstructed each chromosome individually in two sequential steps:

1. Find a set of the most parsimonious arrangements of copy levels on the two parental alleles (**allelic partitioning**).
2. Find the most likely set of SCNA events that would give rise to these copy-number profile (**allele deconstruction**).

Allelic partitioning

Our data consist of integer copy-numbers of each allele at each locus. The data are segmented, with infrequent changes in copy-number between adjacent markers on the array (fewer than one breakpoint per 1000 markers). We start with no information about which copy levels or breakpoints belong on the same allele. The purpose of this section is to find a set of the most parsimonious partitions of copy levels between the two alleles.

There is some information inherent in the structure of the segmentation. Since breakpoints are rare, our model should minimize the number of breakpoints and remove those that are not necessary to explain our observation. These breakpoints are common in situations where a deletion preceded amplification. The first step of our algorithm is to remove such unnecessary breakpoints. Once unnecessary breakpoints are removed, there remains ambiguity about the segmentation of each allele.

There are two situations that lead to ambiguity in segment partitioning between the two alleles: 1) the two alleles are at the exact same copy level at a particular locus, or 2) both alleles have a breakpoint at the exact same SNP marker. The first situation is common, with the second being much rarer. However, in either case we lose the ability to say whether segments preceding that position occurred on the same or opposite allele as segments subsequent to this position. We call these loci “flex-points” as we are free to swap segments between the two alleles only in these regions. We label regions between adjacent flex-points “contigs”, as the partitioning of these segments relative to one another is fixed. The total number of possible arrangements of a given chromosome is 2^{f-1} where f is the number of flex-points on the chromosome.

If there are nine or fewer flex-points on a given chromosome, we enumerate all possible permutations of the contigs across the two alleles (256 different arrangements) and test each of them against each other in event deconstruction (below). This accounted for 98.8% of all chromosomes and removes the complication of allelic ambiguity by brute force. If there are nine or more flex-points, such enumeration is computationally prohibitive, and we focus on the most likely allelic partitions.

To choose a set of partitions to test in deconstruction, we group segments by their copy level and chose a set of allelic phases based on assigning a priority to each set of copy level segments. For each assignment of copy number priority, we took the copy level of highest priority and found the allelic structure that assigned as many segments of that copy level to the same allele. We then fixed these segments and applied the same technique to the copy level of the second priority, recognizing that optimization of the second copy level can be constructed on either allele. Of the 1.2% of chromosomes in which we could not enumerate all potential phasings, 77% had less than 7 unique copy levels (excluding the zero level) and in such cases, we permuted the priority structure of every copy level present on the chromosome. In cases where there were 7 or more unique copy levels, we assumed the most likely partitions would tend to assign unlikely copy-levels (which are rare across the set overall) to the same allele, so that they could be accounted for by a single event rather than requiring separate unlikely events on each allele. In this case, we ranked each copy level based on the log-likelihood of all segments at that copy level across both alleles, assuming each segment is a different event. This created a rank priority structure for each copy level. We then

only permuted the 7 copy levels with the highest priority structure in order to maintain reasonable potential deconstructions.

Allele Deconstruction

Once the segments have been fixed to each allele, SCNA determination is performed in similar fashion to methods described previously[1,75], which identify the combination of SCNAs that would result in the observed copy-number profile and have maximum likelihood of having occurred. The likelihood of an SCNA occurring is estimated according to the observed frequencies of SCNAs with similar lengths and amplitudes of copy-number change across the entire dataset. In contrast to the previous implementation of this algorithm, we consider discrete copy-number values, whereas prior methods focused on continuous total copy ratios. The simplified, discrete data allowed for added complexity when testing the for the MLE solution of chromosome deconstruction. Specifically, the added precision allows confidence about the euploid level (allelic copy level=1 for most samples) as well as loci that have been deleted (allelic copy level =0) and we take advantage of this to search for deconstructions with potentially higher log-likelihoods (LL).

One improvement was directly testing for deletions followed by amplifications over the affected region. For each chromosome, we attempted a deconstruction (as done previously) as well as one where the deleted regions are assessed, then removed prior to chromosome deconstruction (simulating deletion followed by further aneuploidy). Secondly, a necessary heuristic in the previous algorithm created a maximum of two independent sections on the chromosome (a left and a right) and

deconstructing each of these independently. This may lead to an overabundance of telomere bounded events, as events that actually start internal to the chromosome may be end up in a solution that forces them to start on the telomere, followed by a SCNA that reverses the copy level change. To avoid these artifacts, we also tested deconstructions that started internal to the chromosome in the case of a euploid telomere, or if both telomeres are at the sample copy level. To more accurately distinguish SCNA likelihoods, we separated our dataset into two populations based on whether they had undergone whole genome doubling (called by absolute) and ran each set independently.

2. SCNA timing relative to WGD and chromosome duplication

We determined the temporal relations of individual SCNAs to WGD using different approaches for deletions and amplifications.

We considered deletions that involved a change from two copies to zero copies of an allele in WGD samples to have likely occurred prior to WGD. Similarly, deletions that involved a change from two copies to one copy of an allele were considered to have occurred after WGD. Other deletions were left uncalled because of ambiguities introduced by surrounding alterations. When determining timing of genome doubling, we did not include arm level or whole chromosome events, as the events of this size are too common to rule out two sequential events that appear to have the same breakpoints.

Amplifications are more ambiguous than deletions because the extra copies of DNA may end up elsewhere in the genome and be affected by subsequent events in those regions. However, because WGD affects the whole genome simultaneously, we

expect estimates of WGD timing based on amplifications to be similar overall to estimates based on deletions. We called events with an even total copy change as occurring prior to WGD and events with odd copy change as occurring after WGD.

The same metrics were used to determine events before or after chromosome duplication. Again, amplifications are more uncertain than deletions because they may involve disparate regions of the genome.

3. Chromothripsis detection

Chromothripsis results from different mechanisms to most focal events, and has a very different distribution across lineages[254]. We identified chromothripsis events in diploid samples based on three features that are observable in copy-number profiles and which have been associated with chromothripsis previously[174]:

1. A single chromosome exhibits an unexpectedly large number of SCNAs given the observed frequency of SCNAs within the sample.
2. SCNAs on this chromosome tend to be more closely spaced than we would expect by chance.
3. The SCNAs are non-overlapping (because they occurred simultaneously) and lead to copy-number changes of +1 or -1.

Prior estimates of rates of chromothripsis have been complicated by uncertainty as to the absolute numbers of copies of change. In our application of these criteria, we evaluated the absolute allelic copy-number data to identify chromosomes that contained

more non-overlapping SCNAs that involved a single-copy change than we would expect by chance, given the number of SCNAs within the sample and using the binomial distribution. From these chromosomes, we applied the additional criterion that these SCNAs should be more tightly distributed within the chromosome than we would expect given a random selection of non-overlapping SCNAs within our dataset. If this criterion was not met, we applied a recursive algorithm to remove the SCNA furthest from the centroid location of the SCNAs potentially derived from chromothripsis, and recomputed these two statistics.

4. Generation of relative copy-number profiles

The pipeline used to generate relative copy-number estimates is attached as Appendix 1 (a more complete description is to be published separately; Tabak et al, in preparation). In brief, probe-level signal intensities from Affymetrix SNP6 .CEL files were normalized to a uniform brightness across arrays and merged to form intensity values for each probeset using SNPFileCreator, a Java implementation of dChip[255,256]. These intensities were mapped to copy-number levels using Birdseed[145] in the case of SNP markers, and on the basis of experiments with cell lines with varying dosage of X in the case of copy-number markers[1]. Recurrent germline copy-number variations (CNVs) were identified across all DNA samples from normal tissue and markers within these regions (representing ~15% of all markers) were removed from further analysis (Appendix 1). Noise was further reduced by application of Tangent normalization (Appendix 1) followed by Circular Binary Segmentation[141,142]. Quality control metrics were applied at various stages in the pipeline (Appendix 2), resulting in the removal of

data representing 23 cancers out of 4957 primary cancers that had been profiled by SNP6 arrays.

HAPSEG[257] and ABSOLUTE[147], running on FireHose[258], were applied to data from 4870 of these cancers, including both the SNP6 data and, when available, whole-exome sequencing data from the same cancers (1069 samples). Of these, purity and ploidy estimates and genome-wide absolute allelic copy-numbers were called in 3847 cancers (**Table 1**). The 200 acute myeloid leukemia samples were not called by ABSOLUTE because they exhibited copy-number alterations across small fractions of their genomes, resulting in insufficient data for accurate calls by the algorithm.

Author contributions

The author, Steven Schumacher, Gordan Saksena, and Andrew Cherniack were responsible sample curation and quality control, with support from Rameen Beroukhim, Matthew Meyerson, and Gad Getz.

The author, Steven Schumacher, Andrew Cherniack, and Scott Carter were responsible for implementation of the Absolute workflow and manual review through Firehose.

The author was responsible for SCNA deconstruction, WGD timing, and chromothripsis algorithm development and implementation, with advice from Scott Carter, Rameen Beroukhim and Matthew Meyerson.

Results

Cancer purities, ploidies, and rates of copy-number alteration within and across cancer types

We analyzed the copy-number profiles of 4934 primary cancer specimens across 11 cancer types (minimum 136 for bladder cancer; maximum 880 samples for breast

cancer; colon and rectal adenocarcinomas were combined; **Table 1**). In each cancer, we determined copy-numbers at each of 1,559,049 loci relative to the median copy-number genome-wide, using Affymetrix SNP6 arrays and previously described algorithms[20,22,192]. For 3847 cancers, we also determined the purity, ploidy, and absolute allelic copy-number profiles[23-25,147] of the malignant cells using SNP6 array data and, in 1069 cases, matched whole-exome sequencing data (**Table 1).** In the other 1087 cases, purity and ploidy estimates were ambiguous and left uncalled. This included all cases of acute myeloid leukemias [AMLs], which exhibit very few SCNAs.

We then inferred the sequence of somatic copy-alteration (SCNA) events that led to each copy-number profile, using the most parsimonious set of SCNAs that could generate the observed absolute allelic copy-numbers. Using a maximum likelihood approach, we reported the most likely series of SCNAs that led to the copy-number profiles generated by ABSOLUTE for each homologous chromosome (henceforth, "allele"). Each SCNA was characterized by its length, amplitude, genomic position, and, when determinable, allele and the timing of its generation relative to neighboring segments (Methods, **Fig. 1a,b**). We identified a total of 202,244 SCNAs, a median of 39 per cancer sample, comprising six categories: focal SCNAs that were shorter than one chromosome arm (a median of 11 amplifications and 12 deletions per sample); arm-level SCNAs that were chromosome-arm length or longer (a median of three amplifications and five deletions per sample); copy-neutral loss-of-heterozygosity events (cnLOHs), in which one allele had been deleted and the other amplified coextensively (a median of one per sample); and whole-genome duplications (WGDs, in 37% of

Table 1: Dataset information by disease

ABSOLUTE coverage		purity/ploidy called		genome doubled	
count	fraction	count	fraction	count	fraction
109	79.0%	90	83%	56	62%
872	99.1%	745	85%	338	45%
585	99.8%	496	85%	214	43%
563	97.1%	485	86%	52	11%
304	98.1%	270	89%	116	43%
492	99.0%	373	76%	75	20%
0	0.0%				
357	100.0%	292	82%	171	59%
338	98.3%	261	77%	167	64%
565	99.6%	459	81%	245	53%
485	97.4%	376	78%	102	27%
4670	94.2%	**3847**	82%	1536	40%

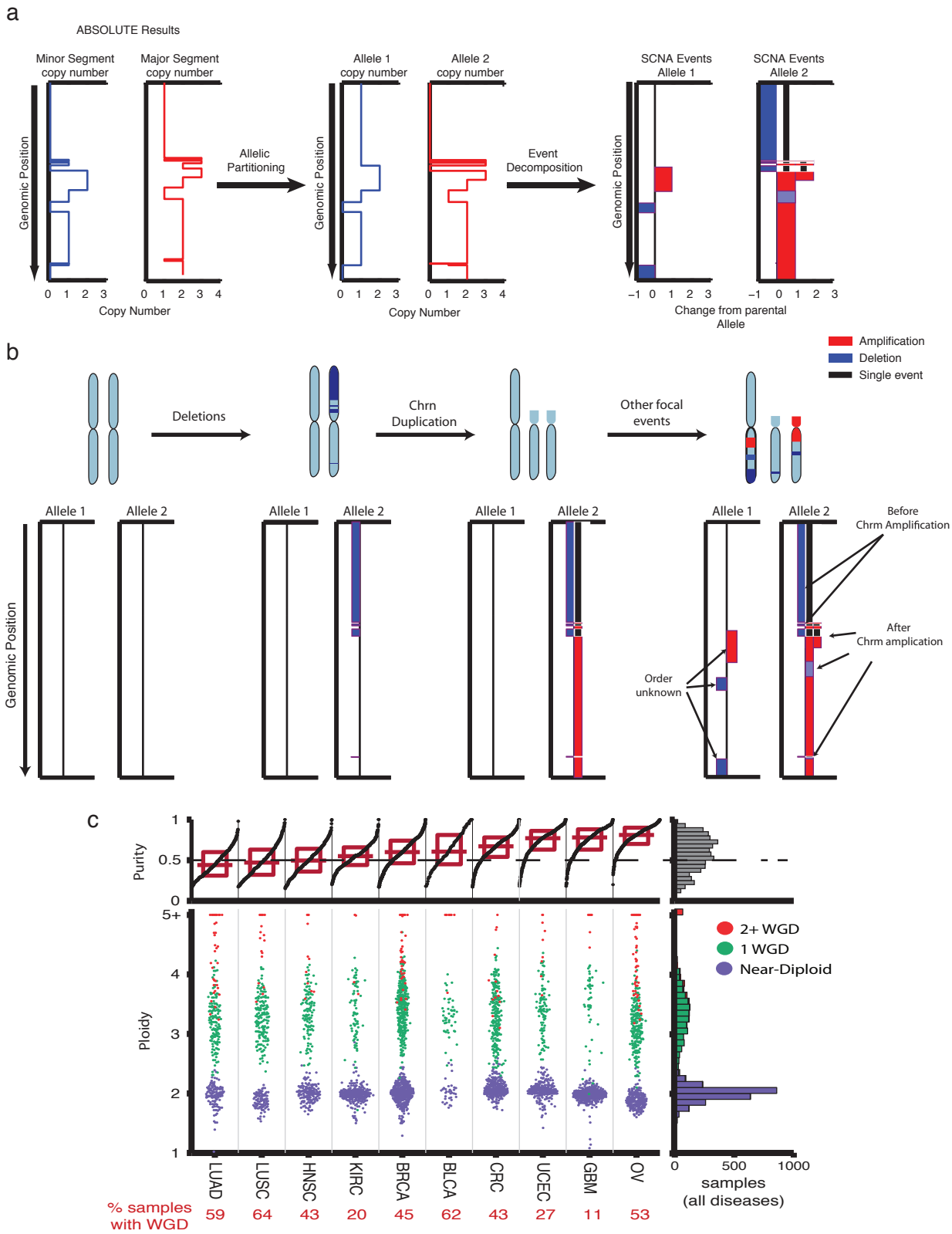

a
ABSOLUTE Results
Minor Segment copy number
Major Segment copy number
Genomic Position
Copy Number
Allelic Partitioning
Allele 1 copy number
Allele 2 copy number
Event Decomposition
SCNA Events Allele 1
SCNA Events Allele 2
Change from parental Allele
b
Amplification
Deletion
Single event
Deletions
Chrn Duplication
Other focal events
Allele 1
Allele 2
Before Chrm Amplification
After Chrm amplication
Order unknown
c
Purity
Ploidy
2+ WGD
1 WGD
Near-Diploid
samples (all diseases)
LUAD
LUSC
HNSC
KIRC
BRCA
BLCA
CRC
UCEC
GBM
OV
% samples with WGD
59
64
43
20
45
62
43
27
11
53

Figure 1:
(**a**) Schematic indicating procedure to determine SCNAs. Absolute allelic copy-numbers generated by ABSOLUTE (left panel) are partitioned with the lowest copy-numbers on one allele (blue) and the higher copy-numbers on the other (red). We repartition the copy-numbers between alleles to test all possibilities if computationally efficient, or the most likely possibilities if not (middle panel). The most likely set of SCNAs generating each copy-number profile is then determined (right panel). Regions in red are amplifications; regions in blue are deletions. The black lines in allele 2 indicate deletions that preceded an amplification, so that segments that appear discontinuous on the reference genome were amplified in a single event. (**b**) Schematic indicating temporal order of SCNAs determined in (a). In this example, deletions followed by a chromosome level amplification can account for the copy-number profiles in (a). For the same profile to be generated by amplifications followed by deletions, either the deletions would have to remove two copy-levels, requiring sequential deletions with identical boundaries, or multiple neighboring, non-contiguous amplifications would be required in addition to multiple, non-contiguous deletions. We assume both of these possibilities are unlikely. (**c**) Sample purities (top panel) and ploidies (bottom panel) across lineages (see Supplementary Table 1 for a list of lineage abbreviations). Near-diploid samples are designated in purple; cancers that have undergone one or more than one WGD event are designated by green and red, respectively. Summarized data across all lineages are indicated on the right.

cancers). By amplifications and deletions, we refer to copy-number gains and losses, respectively, of any length and amplitude.

Estimated purities and ploidies per cancer varied substantially within and across diseases (**Fig. 1c)**. The purity estimates correlated with estimates derived from measurements of leukocyte and lymphocyte contamination using DNA methylation data from the same cancers (**Fig. 2a**) (Shen et al, unpublished data), but tended to indicate lower purity, consistent with the presence of non-hematopoietic contaminating normal cells.

Average ploidies within diseases mirrored their frequencies of WGD (**Fig. 1c**). The average estimated ploidy within samples that had undergone a single WGD was 3.31 (not four), suggesting that WGD events are associated with large amounts of genome loss. By contrast, samples that had not undergone WGD had an average estimated ploidy of 1.99.

Compared to the near-diploid cancers within each disease, cancers with WGD had higher rates of every other type of SCNA (**Fig. 2b**) and twice the rate of SCNAs overall. Across diseases, overall SCNA rates largely reflected rates of WGD (**2c**).

In cancers with WGD, most other SCNAs occurred after WGD (**Fig. 2d, see Methods**). The fractions of amplifications and deletions that were estimated to occur prior to WGD were highly correlated across diseases (R=0.64, **Fig. 2d**), indicating a consistent estimate for the timing of WGD with respect to other SCNAs. WGD was inferred to occur earliest relative to focal SCNAs among diseases where WGD was common (ovarian, bladder, and colorectal cancers), and after most focal SCNAs in

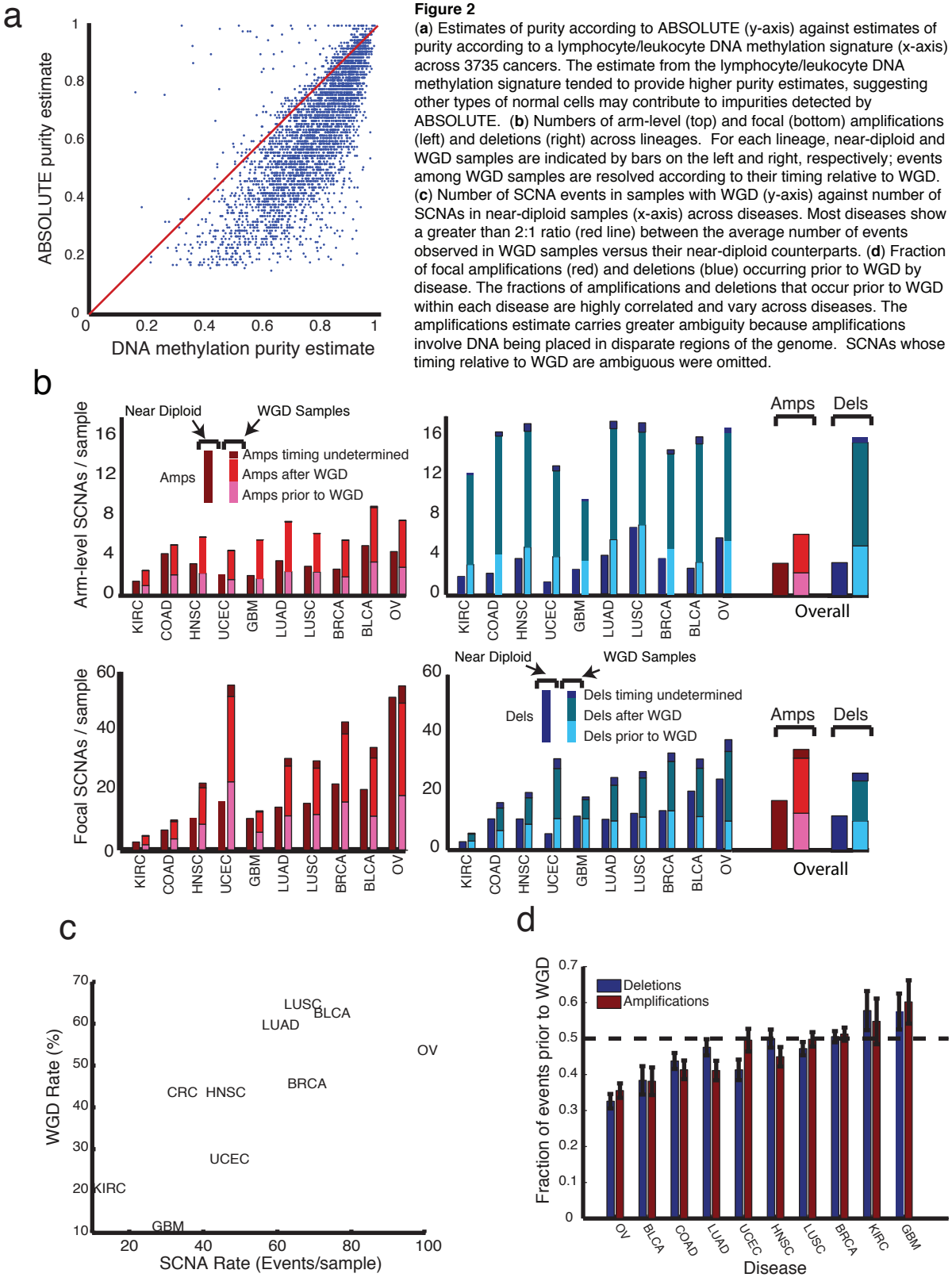


Figure 2

(**a**) Estimates of purity according to ABSOLUTE (y-axis) against estimates of purity according to a lymphocyte/leukocyte DNA methylation signature (x-axis) across 3735 cancers. The estimate from the lymphocyte/leukocyte DNA methylation signature tended to provide higher purity estimates, suggesting other types of normal cells may contribute to impurities detected by ABSOLUTE. (**b**) Numbers of arm-level (top) and focal (bottom) amplifications (left) and deletions (right) across lineages. For each lineage, near-diploid and WGD samples are indicated by bars on the left and right, respectively; events among WGD samples are resolved according to their timing relative to WGD. (**c**) Number of SCNA events in samples with WGD (y-axis) against number of SCNAs in near-diploid samples (x-axis) across diseases. Most diseases show a greater than 2:1 ratio (red line) between the average number of events observed in WGD samples versus their near-diploid counterparts. (**d**) Fraction of focal amplifications (red) and deletions (blue) occurring prior to WGD by disease. The fractions of amplifications and deletions that occur prior to WGD within each disease are highly correlated and vary across diseases. The amplifications estimate carries greater ambiguity because amplifications involve DNA being placed in disparate regions of the genome. SCNAs whose timing relative to WGD are ambiguous were omitted.

diseases in which WGD was least common (glioblastoma and kidney clear cell carcinoma).

SCNA lengths suggest varied mechanisms of generation

Focal SCNAs for which one boundary is the telomere (telomere-bounded) tend to be longer than SCNAs in which both boundaries are internal to a chromosome (median SCNA length: amplifications 19.6 Mb versus 0.9 Mb; deletions: 22.7 Mb versus 0.7 Mb, for telomere-bounded and internal events respectively). These differences reflect differences across the entire length distributions of internal and telomere-bounded events. Focal internal SCNAs were observed at frequencies inversely proportional to their lengths (**Fig. 3a-b**), as noted previously[26]. However, telomere-bounded SCNAs tend to follow a superposition of 1/length and uniform length distributions. These distributions are the same whether measuring distance by kb, number of array markers, or number of genes, indicating that they do not result from variations in array resolution or gene density genome-wide (data not shown). Focal, telomere-bounded SCNAs also accounted for more SCNAs (12% and 26% of focal amplifications and deletions, respectively) than expected assuming random SCNA locations ($p<0.0001$). Both telomere-bounded and internal SCNAs are more likely to end within the centromere than expected given the centromere's length (**Fig. 3c)**, but the differences in their length distributions remain when centromere-bounded events are excluded. Differences between telomere-bounded and internal SCNAs are even more marked for cnLOH events (**Fig. 3d)**.

We detected chromothripsis in 5% of samples, ranging from none of head and neck squamous cell carcinomas to 16% of glioblastomas (**Fig. 4a; see Methods**). The

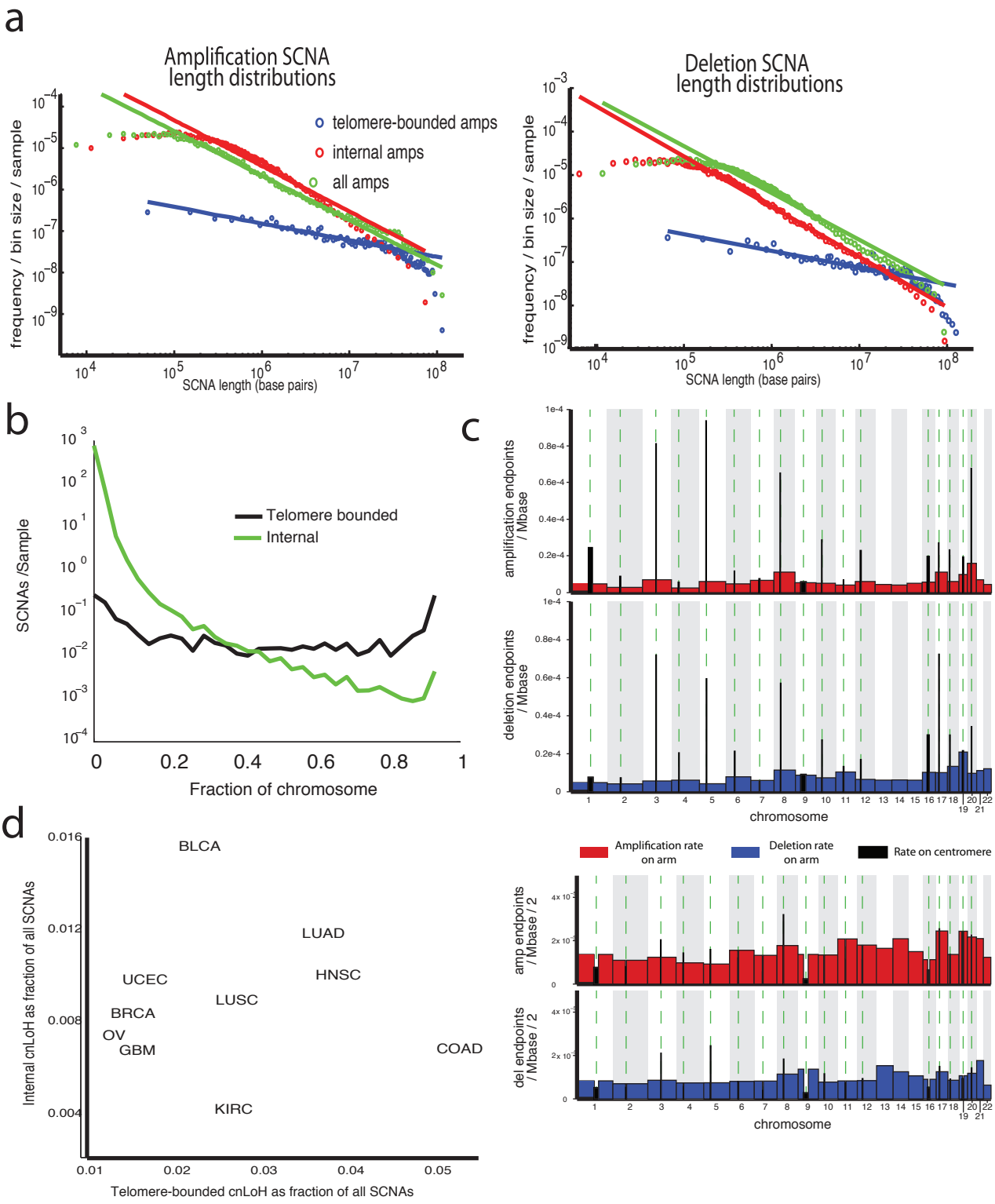

a
Amplification SCNA length distributions
telomere-bounded amps
internal amps
all amps
frequency / bin size / sample
SCNA length (base pairs)
Deletion SCNA length distributions
frequency / bin size / sample
SCNA length (base pairs)
b
Telomere bounded
Internal
SCNAs /Sample
Fraction of chromosome
c
amplification endpoints / Mbase
deletion endpoints / Mbase
chromosome
Amplification rate on arm
Deletion rate on arm
Rate on centromere
amp endpoints / Mbase / 2
del endpoints / Mbase / 2
chromosome
d
BLCA
LUAD
HNSC
UCEC
LUSC
BRCA
OV
GBM
COAD
KIRC
Internal cnLoH as fraction of all SCNAs
Telomere-bounded cnLoH as fraction of all SCNAs

Figure 3: SCNA length distribution

(**a**) Frequencies of all amplifications (green) and amplifications that begin at the telomere (blue) or are internal to the chromosome (red) against amplification length. We used a fixed number of events per bin as opposed to a fixed bin size to compute our least squared fit because there were many fewer long events. The frequencies of amplifications that are greater than 400 kb and less than a chromosome arm in length follow a power law $f(L) = 1/L^{\beta}$, where for all events, $\beta = 1.05$ and $r^2=0.99$; for telomeric events $\beta = 0.45$ and $r^2=0.90$; and for internal events, $\beta = 1.12$ and $r^2=0.97$ (**b**) The distribution of lengths of SCNAs originating at telomeres (black line) compared to SCNAs that are internal to the chromosome. **(c)** Number of amplification (red, top) and deletion (blue, bottom) endpoints in chromosome arms (colored regions) and centromeres in metacentric chromosomes (black regions) for telomere-bounded SCNAs (left panels) and internal SCNAs (right panels). SCNAs are more likely to end within the centromere than expected given the centromere's length for both telomeric and internal SCNAs ($p<0.0001$ in both cases). In the case of telomeric events, this tendency results in a propensity for SCNAs to be arm-level events involving precisely one chromosome arm and suggests that focal telomeric SCNAs are generated by similar mechanisms to such arm-level events. Note that centromeres often span large regions without SNP array markers, preventing detection of many SCNA endpoints in these regions. The width of each region reflects the size of its genomic locus. (**d**) Rates of cnLOH events that were internal to chromosomes (y-axis) against rates of cnLOH events that were telomere-bounded (x-axis) across diseases. Only 2% of focal SCNAs were cnLOH, and these events had more pronounced differences between telomeric and internal events than did amplifications and deletions. Most cnLOHs (58%) involved either whole chromosomes or exactly one chromosome arm, compared to an 18% rate of arm- and chromosome-level events for other SCNAs ($p < 0.0001$). Internal cnLOHs were typically much smaller than other focal internal SCNAs (median 0.2 Mb vs 0.8 Mb, Mann-Whitney (MW) $p < 0.0001$), whereas telomeric cnLOHs were much larger than other telomeric SCNAs (median 82Mb vs 27.2Mb, $p<0.0001$). Rates of telomeric and internal cnLOH show no correlation across diseases, suggesting the processes that lead to these events may be distinct.

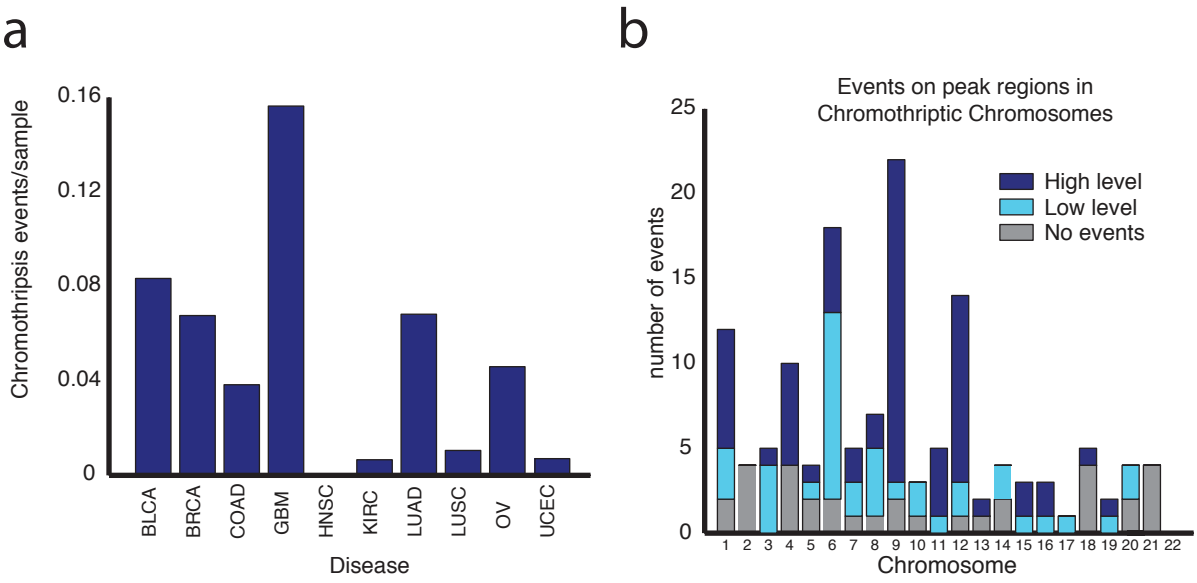


Figure 4-Chromothripsis across cancer
(**a**) Rates of chromothripsis across lineages. (**b**) Rates of chromothripsis across chromosomes. Chromothripsis events that involved peak regions of amplification and deletion (see below) are indicated in blue (dark blue: amplifications >4.4 copies or deletions<-1; light blue: low-level events involving smaller changes); events that do not involve peak regions are indicated in grey.

rate of chromothripsis was not related to overall rates of SCNA (r=0.13, p=0.3). As previously reported[259], samples with chromothripsis were more likely to have chromothripsis on more than one chromosome (14/122 samples with chromothripsis had two to three such events, p=0.003). Many chromothripsis events were concentrated in a few genomic regions, often associated with known driver events (**Fig. 4b**). In glioblastomas, chromothripsis events were concentrated in chromosomes 9 and 12 and corresponded respectively to homozygous loss of CDKN2A (20/22 samples) and coamplification of discontinuous regions containing CDK4 and MDM2 (9/12 samples). Across all cancers, 72% of chromothripsis events included a GISTIC peak region (see below).

Discussion

Comprehensive analysis of SCNA across human tumors

This study represents the largest analysis to date of high-resolution copy-number profiles generated using a single platform, and the largest analysis of absolute allelic copy-number data across cancer types. The TCGA Pancancer project was designed to combine the efforts of many individuals and institutions and comprehensively characterize a large, well annotated collection of primary tumors. Cataloging and analyzing the SCNAs in this set is an integral part of this endeavor and will be a foundation for future studies such as those described in Chapter 3.

Assessing sample purity and its importance on primary tumor analysis

As current genetic analysis techniques involve interpretation of bulk tumor samples, purity of the tumor within the biopsy, and overall ploidy of the cancer cells can confound both detection of SCNAs and significance analyses. Even worse, driver

alterations, such as BRAF-V600E[260] status or *TP53* inactivation[261] may correlate with purity or ploidy, biasing the significance we assign to important steps in cancer development. Besides allowing us to reject impure samples, we can build models that correct for these confounders both in current and future significance studies. We found that purity and ploidy varied greatly both within and across disease, with lung and head and neck cancers being the most impure tumors, and GBM and ovarian being relatively pure. Outside of SCNA analysis, these purity values should be utilized in other analyses of TCGA data, including expression, DNA methylation, and mutational analyses.

Selective pressures.

A primary challenge in the analysis of somatic genetic data is distinguishing between patterns of alteration that reflect the mechanism by which those alterations are generated, from those that represent selection, specifically positive selection. An underlying assumption of our current analyses is that patterns of alteration that are observed across all chromosomes are likely to reflect mechanistic biases, whereas deviations from these patterns at individual loci are likely to reflect selective pressures. However, cancer is a complicated evolutionary process and improving our understanding of the background rate of alteration will improve current and future significance models.

We identified common patterns of SCNA across cancer types, including a tendency for telomeric events to be longer and more frequent than SCNAs within chromosomes. The differences between telomere-bounded and internal SCNAs across all chromosomes suggest different mechanisms underlie their generation. Internal SCNAs have been proposed to occur as a result of apposition of their two breakpoints

in three-dimensional space. Chromatin is arranged as a “fractal globule” during interphase[262,263], during which time the likelihood that two breakpoints would be apposed decreases proportionally to the linear distance between them, implying a 1/length distribution. Conversely, SCNAs that start on the telomere may be related to telomere shortening and telomere crisis, and associated with a single double-strand break that could occur anywhere within the chromosome[264]. These, and other[185,265], mechanistic disparities should be accounted for in the background model of SCNA significance analyses and we are currently in the process of designing new significance tests that have the flexibility to do so.

Negative selection is another factor that significantly shapes the cancer genome. In our data, for example, homozygous deletions was detected in around 0.04% of markers, far below what is expected rate (0.2%, excluding the possibility of homozygous arm-level loss). The inability to tolerate the deletion of essential genes can directly affect significance analysis by highlighting regions unaffected by negative selection (such as gene poor regions), as potentially important. For example, proximity to an essential gene may alter the observed events leading to tumor suppressor inactivation. Finally, events during cancer evolution may increase tolerance for deletion, such as our finding that duplications of large regions of the genome (through WGD or polysomy) tend to lead to subsequent increases in numbers of SCNAs (especially deletions) in the duplicated regions. Moving forward, whole-genome sequencing data can indicate the specific rearrangements that contributed to each SCNA[266,267], and assessment of genetic heterogeneity within tumors can also distinguish early from late events[147,268].

Both of these are approaches are likely to inform the mechanisms by which SCNAs are generated and the selective pressures that shape them.

Chromothripsis is a relatively new mechanism for somatic copy number change, and its mechanism is still not fully understood. The large-scale structural rearrangements present in this event make whole genome sequencing the ideal technique for their identification. However, we currently have many thousands more primary tumors analyzed by microarray technology than by WGS, so an algorithm that can identify patterns of chromothripsis without the aid of sequencing technology would be useful. We found rates of chromothripsis did not track with rates of other SCNAs, with rates specifically higher in GBM. Further refinement of this technique in WGD samples may allow us to determine whether rates of chromothripsis change after WGD.

Chapter 3: Significance of recurrent, focal SCNAs across disease

Section Goals

A primary challenge in understanding SCNAs is to distinguish the driver events that contribute to oncogenesis and cancer progression from the passenger SCNAs that are acquired during cancer evolution but do not contribute towards it[50,190,269,270]. Positively selected SCNAs will tend to recur across cancers at elevated rates[191,254,271]. However, SCNAs may also recur in the absence of positive selection due to increased rates of generation or decreased negative selection[272,273]. For this reason, it is important to understand how mechanisms of SCNA generation, their temporal ordering, and negative selection shape the distribution of SCNAs genome-wide[268,272,273].

A second challenge is to identify the oncogene and tumor suppressor gene targets of the driver SCNAs (which often encompass many genes) and elucidate the SCNA's functional roles. The context of the SCNA can be informative. Positive correlations with other genetic events may indicate functional synergies, while anticorrelations may indicate functional redundancies because redundant events would not be required by the same cancer. Several approaches have been developed to determine functional effects of genetic events based on anticorrelation patterns[200,202,274].

Here, we address these challenges through analyses of the 4934 cancer copy-number profiles described in Chapter 2. These profiles, spanning 11 cancer types and assembled through The Cancer Genome Atlas Project Pan-Cancer effort, enable sensitive determination of significant regions of copy-number change both within and

across cancer types. We have also developed new approaches to identify functionally relevant correlations between SCNAs. Among these are correlates with WGD, including *TP53* mutations, *CCNE1* amplifications, and alterations of the PPP2R complex. Finally, we used SCNA timing relative to WGD to identify regions of LOH that occur early in tumor evolution, allowing us to separate potential initiating regions of significant deletion from those that may be the result of increased rate of generation.

Methods

1. Impurity-corrected GISTIC

The less pure a sample, the less signal is derived from the somatic alterations present in the tumor, which will decrease its contribution in recurrence analyses. In cases where we were able to estimate purity and ploidy from ABSOLUTE, we "corrected" total copy-ratios for signal dampening due to cancer cell impurity (i.e. contamination with normal DNA). We called this In-Silico Admixture Removal (ISAR). The observed copy-ratio R(x) at locus x is a function of the purity α, cancer cell ploidy τ (representing the average copy-number genome-wide), and integer copy-number (in the cancer cells) q(x)[147]

$$R(x) = (\alpha \; q(x) + 2 (1- \alpha)) / D ,$$

where D represents the average ploidy across all cells in the cancer:

$$D = \alpha \tau + 2 (1 - \alpha).$$

From this, we can determine q(x):

$$q(x) = D\,R(x) / \alpha - 2\,(1 - \alpha) / \alpha\ .$$

We assume that the functionally relevant number is the copy-ratio within cancer cells, representing the integer number of copies q(x) divided by the overall ploidy of the cell τ:

$$R'(x) = q(x) / \tau = R(x) / \alpha - 2\,(1 - \alpha) / (\alpha\ \tau)\ .$$

Use of R'(x) has the effect of amplifying the signal from low purity samples to be equivalent to higher purity samples. For samples for which ABSOLUTE calls were not available, we used R(x).

To determine significantly recurrent regions of SCNA, we used GISTIC 2.0[75] applied to the transformed copy-number data. We used a noise threshold of 0.3, a broad length cutoff of 0.5 chromosome arms, a confidence level of 95%, and a copy-ratio cap of 1.5.

For some lineage-specific analyses, dozens of regions on a single chromosome arm were identified as significant peaks due to the presence in many samples of discontinuous SCNAs (such as chromothripsis) on those chromosome arms. This phenomenon has been observed previously[271]. We narrowed these regions by applying in all lineage-specific analyses an "arm-level peel-off" correction that considers all SCNAs on a chromosome arm in a single sample to be part of a single event when

determining whether multiple significantly recurrent events exist on that chromosome arm. This approach has also been used in prior analyses[275].

The genes listed in each peak region include all protein-coding genes and microRNAs and additional non-coding RNAs as listed in the files refGene.txt, refLink.txt, refSeqStatus.txt, and wgRna.txt from the UCSC Golden Path database (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/) as of 27 February 2012.

2. Significance of chromatin modifying genes among peak regions of amplification without known driver genes

To determine whether epigenetic regulators were enriched in peak regions, we compared the number of regions with epigenetic regulators (using a published list[276]) to permuted datasets in which each gene in each region was replaced by a gene randomly selected from elsewhere in the genome.

3. Correlation analysis

To determine the significance of correlations and anticorrelations between pairs of SCNAs, we compared the observed rate of co-occurrences to the rate of co-occurrences in 5000 permuted copy-number profiles for which we had randomized the sample assignment for each chromosome, while maintaining genomic position and lineage and sub-lineage assignments. We only considered SCNAs in different chromosomes to avoid confounding due to geographic proximity.

To control for variable rates of genomic disruption across samples, we modified the permutations so that they maintained both the numbers of amplified and deleted

markers A^0_j and D^0_j in each sample *j*. After randomizing sample assignments for each chromosome as described above, we applied simulated annealing[277,278] in which we picked a chromosome at random and swapped it between two randomly chosen samples within the same lineage at each step, and accepted the step with a probability 1- E_{tot}, where:

$$E_{tot} = T_{amp} * \sum_j \frac{(A_j^{t+1} - A_j^0)}{A_j^0 + 1} + T_{del} * \sum_j \frac{(D_j^{t+1} - D_j^0)}{D_j^0 + 1}$$

and A^t_j and D^t_j represent the numbers of amplified and deleted markers in sample *j* and step *t*. T_{amp} and T_{del} are temperature factors that were slowly increased during the annealing, and the 1 in the denominator of each value is to avoid dividing by 0 in samples without any events. This procedure was applied in two separate analyses: one in which we looked at all SCNAs that passed the noise thresholds we used for our GISTIC significance analyses (above), and one in which we only considered loci with copy-number <-1 or >4.4. The second analysis we termed our "high-level" analysis.

4. Intersection between mutual exclusivity network and Dapple network

To validate the functionality of our network, we looked at the overlap between our network and DAPPLE, a curated dataset of protein-protein interactions[279] (PPIs). Of the >400,000 PPI pairs, we took only pairs with a score equal to 1 (indicating highest confidence). Two peak regions had an edge between them in the PPI network under two conditions;

1. A protein within the first peak was a direct interactor with a protein in the second peak.
2. A protein in the first peak had at least three distinct paths of length 2 in the PPI network to a protein in the second peak.

To improve specificity, we only tested regions containing fewer than 25 genes. We determined whether the similarity between the PPI network and the anticorrelation network was significant by comparing the extent of overlap to permutations in which the edges in the anticorrelation network were randomly reassigned while maintaining the overall connectivity of the graph (see Results). By comparing both observed and anticorrelation networks to the same PPI network, we controlled for the propensity of regions with many genes to map to more PPIs.

5. Somatic genetic correlates with WGD

To determine which of the 200 most significant somatic mutations correlate with WGD, we used the *permmatswap* function in the R[280] package "vegan"[281] with the "quasifit" handle[34] to produce a series of independent assignments for mutations on each gene within each sample. This function maintained the number of mutations per gene per lineage, as well as the number of the number of mutations per sample.

To determine which of the peak regions had SCNAs that correlate with WGD, we compared the number of times each SCNA was observed in WGD samples in our observed data to the number of times the SCNA was observed in WGD samples in the permutations created by our simulated annealing approach above.

6. Overlap of peak regions of SCNA

Two regions were considered to overlap if their 95% confidence intervals intersected. To determine significance of overlap, we compared the number of peak regions that overlapped across at least two lineages in the observed data to 100,000 permutations in which the locations of each peak region were randomly shuffled within its chromosome arm (disallowing extension past the telomere or centromere).

7. GRAIL analysis

We used GRAIL[282] (www.broadinstitute.org/mpg/grail/) to find common functional terms in the literature for the genes in peak regions of SCNA. We used only PubMed abstracts through December 2006. We removed the following non-informative keywords from those GRAIL found most significant: "growth", "cancer", "cancers", "tumor", "tumors", "proliferation", "suppressor", "factors", "loss", "like", "rich", "cel", "cells", "yeast", "system", "family", "repeat", "deletions", "elegans", "national".

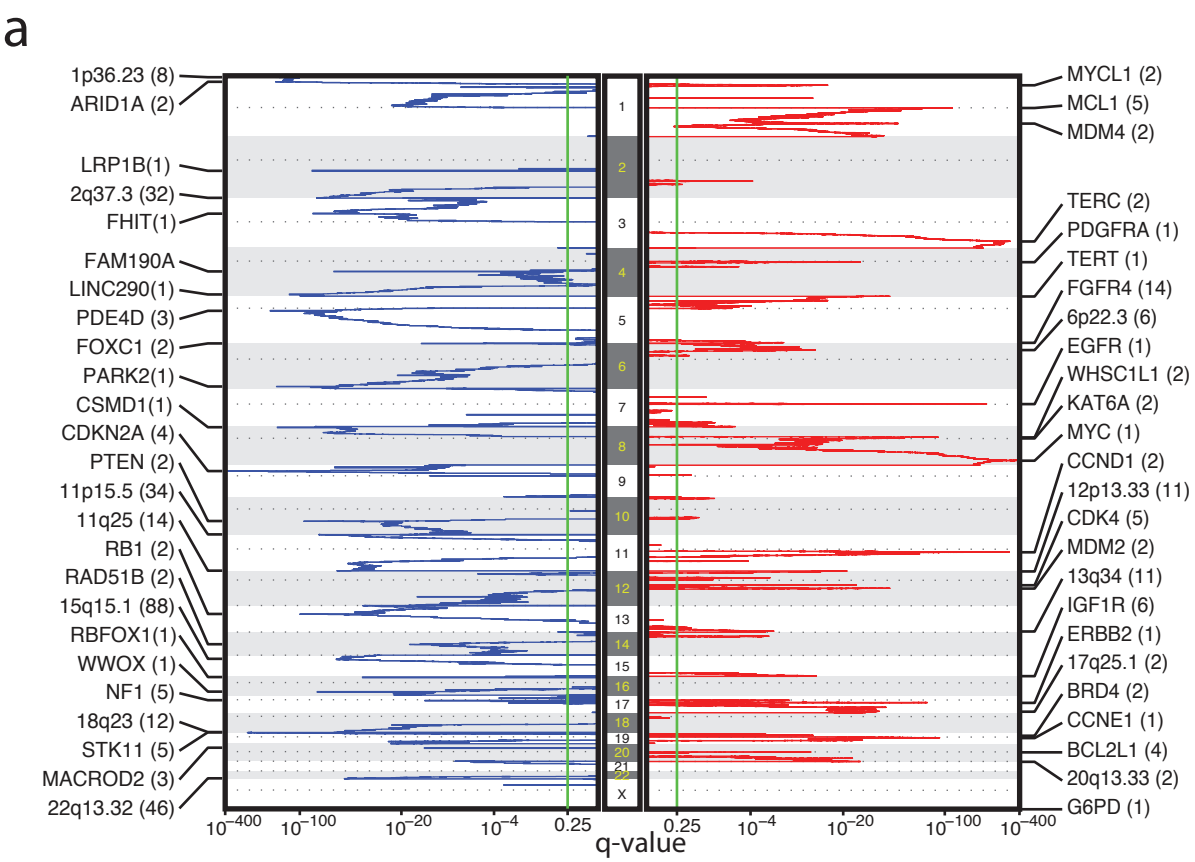
Author Contributions

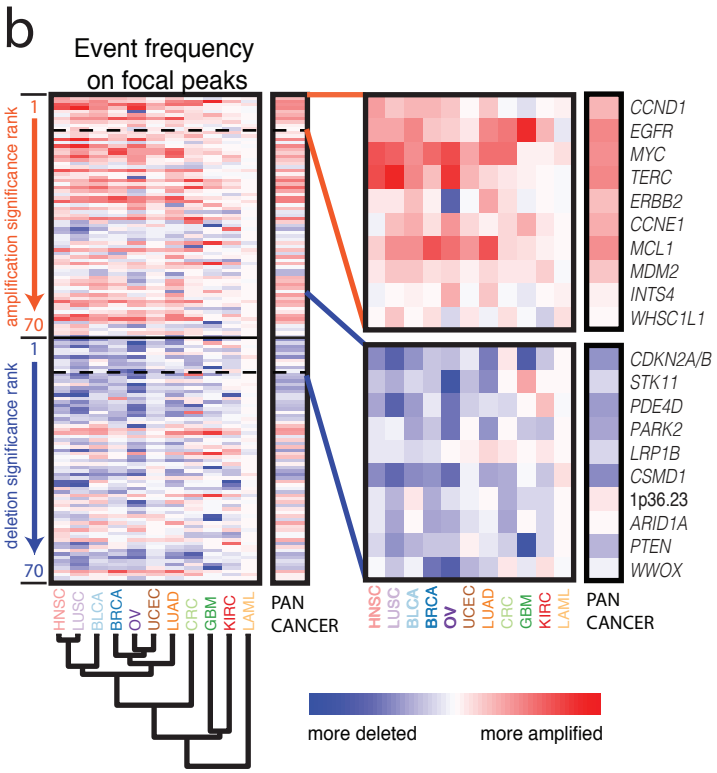
The author was responsible for event and significance analyses of relative data and development of Correlation analysis with support and advice from Steven Schumacher, Rameen Beroukhim, Matthew Meyerson, and Gad Getz.

Results

Significant regions of Somatic Copy Number Alteration

We identified 70 recurrently amplified and 70 recurrently deleted regions in a unified “Pan-Cancer” analysis across all lineages (**Fig. 5a, Table 2**). SCNAs involving





c

Associated terms in literature

All peak regions

Amplifications	Deletions
1. Cyclin	1. PTEN
2. Telomerase	2. Phosphatase
3. Transcription	3. Prostate
4. Expression	4. Mutations

Peak regions without well documented drivers

Amplifications	Deletions
1. Histone	1. Phosphatase
2. Cytochrome	2. RNAI
3. Mitochondrial	3. PTEN
4. Acetyltransferase	4. Prostate

Figure 5- Pancancer significance analysis
(**a**) Significance of focal SCNAs. GISTIC q-values (x-axis) for deletions (left, blue) and amplifications (right, red) are plotted across the genome (y-axis). Candidate gene targets within each peak are indicated for the 25 most significant peaks; in cases where no clear candidate was identified, the cytoband was indicated. Values in parentheses indicate the number of genes in each peak. Green lines indicate the significance threshold (q=0.25). (**b**) Frequencies of amplification minus frequencies of deletion (red and blue indicated propensity to amplifications and deletions, respectively) across lineages (x-axis; see Supplementary Table 1 for a list of lineage abbreviations) for all 84 significant peak regions of SCNA, arranged in order of significance (y-axis). The ordering of lineages reflects the results of unsupervised hierarchical clustering of these data. Magnified views of the values for the ten most significant amplification and deletion peaks, respectively, are shown to the right, alongside candidate targets for these regions. Criteria for selecting the indicated candidates are described in the Methods. (**c**) Associated terms in literature in peak regions containing fewer than 25 genes, according to a GRAIL analysis of (top) all peak regions and (bottom) peak regions without known cancer genes or large genes.

Table 2: Pan-cancer regions of significant SCNA

A) Amplification

Peak Name	Rank	Genomic location	Peak region	GISTIC q-value	Gene count	Target(s)	Frequently mutated genes[B]
CCND1	1	11q13.3	chr11:69464719-69502928	2.05E-278	2	*CCND1*[K]	*CCND1* = 6.6e-08
EGFR	2	7p11.2	chr7:55075808-55093954	2.30E-240	1	*EGFR*[K]	*EGFR* = 2.2e-15
MYC	3	8q24.21	chr8:128739772-128762863	6.50E-180	1	*MYC*[K]	
TERC	4	3q26.2	chr3:169389459-169490555	5.40E-117	2	*TERC*[P]	
ERBB2	5	17q12	chr17:37848534-37877201	1.59E-107	1	*ERBB2*[K]	*ERBB2* = 1.3e-06
CCNE1	6	19q12	chr19:30306758-30316875	4.77E-90	1	*CCNE1*[K]	
MCL1	7	1q21.3	chr1:150496857-150678056	1.25E-80	6	*MCL1*[K]	
MDM2	8	12q15	chr12:69183279-69260755	2.59E-62	2	*MDM2*[K]	
INTS4	9	11q14.1	chr11:77610143-77641464	1.01E-54	1	*INTS4*	
WHSC1L1	10	8p11.23	chr8:38191804-38260814	3.43E-46	2	*WHSC1L1*[E], *LETM2*[M]	
CDK4	11	12q14.1	chr12:58135797-58156509	5.14E-41	5	*CDK4*[K]	*CDK4* = 0.0048
KAT6A	12	8p11.21	chr8:41751300-41897859	2.97E-39	2	*KAT6A*[P,E], *IKBKB*[**]	
SOX2	13	3q26.33	chr3:181151312-181928394	1.21E-38	2	*SOX2*[K]	
PDGFRA	14	4q12	chr4:54924794-55218386	1.08E-37	3	*PDGFRA*[K]	
BDH1	15	3q29	chr3:197212101-197335320	1.21E-31	1	*BDH1*[M]	
1q44	16	1q44[T]	chr1:242979907-249250621	4.48E-31	83		
MDM4	17	1q32.1	chr1:204367383-204548517	1.98E-29	3	*MDM4*[K]	
TERT	18	5p15.33	chr5:1287704-1300024	9.34E-27	1	*TERT*[K]	
KDM5A	19	12p13.33[T]	chr12:1-980639	1.59E-25	11	*KDM5A*[E]	
MYCL1	20	1p34.2	chr1:40317971-40417342	3.99E-25	2	*MYCL1*[K]	
IGF1R	21	15q26.3	chr15:98667475-100292401	8.62E-25	9	*IGF1R*[K]	
PARP10	22	8q24.3	chr8:144925436-145219779	5.44E-20	15	*PARP10*[P,E], *CYC1*[M]	
G6PD	23	Xq28	chrX:153760870-153767853	3.66E-19	1	*G6PD*	
PHF12	24	17q11.2	chr17:27032828-27327946	1.75E-16	21	*PHF12*[E], *ERAL1*[M]	
20q13.33	25	20q13.33	chr20:62187847-62214354	2.96E-16	2		
PAF1	26	19q13.2	chr19:39699366-39945515	1.66E-15	13	*PAF1*[P,E]	*IL28A*=0.021, *SUPT5H*=0.084
BCL2L1	27	20q11.21	chr20:30179028-30320705	2.85E-15	4	*BCL2L1*[K]	
TUBD1	28	17q23.1	chr17:57922443-57946458	7.19E-15	1	*TUBD1*	*TUBD1* = 0.009
[ZNF703]	29	8p11.23	chr8:37492669-37527108	2.44E-14	0		
1q23.3	30	1q23.3	chr1:160949115-161115281	7.73E-13	9		
8q22.2	31	8q22.2	chr8:101324079-101652657	4.22E-11	3		*SNX31* = 0.015
BRD4	32	19p13.12	chr19:15310246-15428182	5.04E-10	3	*NOTCH3*[P], *BRD4*[P,E]	
KRAS	33	12p12.1	chr12:24880663-25722878	9.47E-10	7	*KRAS*[K]	*KRAS* = 1.5e-14
NKX2-1	34	14q13.2	chr14:35587755-37523513	1.33E-09	14	*NKX2-1*[K]	*NFKBIA*=0.0098, *RALGAPA1*=0.027
NFE2L2	35	2q31.2	chr2:178072322-178171101	5.48E-09	5	*NFE2L2*	*NFE2L2* = 3.9e-14
ZNF217	36	20q13.2	chr20:52148496-52442225	5.83E-08	1	*ZNF217*[K]	*ZNF217* = 0.0082
13q34	37	13q34[T]	chr13:108818892-115169878	6.28E-08	45		*ING1* = 0.00026
KAT6B	38	10q22.2	chr10:76497097-77194071	1.41E-07	9	*KAT6B*[E], *VDAC2*[M]	
NSD1	39	5q35.3	chr5:176337344-177040112	1.75E-06	22	*NSD1*[E], *PRELID1*[M]	*NSD1* = 4.9e-10
FGFR3	40	4p16.3	chr4:1778797-1817427	2.14E-06	2	*FGFR3*[P], *LETM1*[M]	*FGFR3* = 0.00018
9p13.3	41	9p13.3	chr9:35652385-35739486	2.55E-06	8		
COX18	42	4q13.3	chr4:73530210-74658151	2.68E-06	7	*COX18*[M]	
7q36.3	43	7q36.3[T]	chr7:153768037-159138663	3.19E-06	30	*PTPRN2*[L], *DPP6*[L]	
18q11.2	44	18q11.2	chr18:23857484-24119078	3.83E-06	2		
SOX17	45	8q11.23	chr8:55069781-55384342	2.02E-05	1	*SOX17*	*SOX17* = 0.00092
11q22.2	46	11q22.2	chr11:102295593-102512085	0.00015337	3		
CBX8	47	17q25.3	chr17:77770110-77795534	0.00023029	1	*CBX8*[E]	
AKT1	48	14q32.33	chr14:105182581-105333748	0.00028451	7	*AKT1*[K]	*AKT1* = 1.1e-14
CDK6	49	7q21.2	chr7:92196092-92530348	0.00069831	3	*CDK6*[K]	
6p21.1	50	6p21.1	chr6:41519930-44297771	0.0010459	70		
EHF	51	11p13	chr11:34574296-34857324	0.0011002	1	*EHF*	
6q21	52	6q21	chr6:107098934-107359899	0.0011806	4		
19q13.42	53	19q13.42[T]	chr19:55524376-59128983	0.0013319	138		*ZNF471* = 5.4e-05
17q21.33	54	17q21.33	chr17:47346425-47509605	0.0025775	2		
BPTF	55	17q24.2	chr17:65678858-66288612	0.0028375	11	*BPTF*[E]	
E2F3	56	6p22.3	chr6:19610794-22191922	0.0033658	7	*E2F3*[K]	
19p13.2	57	19p13.2	chr19:10260457-10467501	0.0038041	12	*MRPL4*[M]	*DNMT1* = 0.099
17q25.1	58	17q25.1	chr17:73568926-73594884	0.012337	2		
KDM2A	59	11q13.2	chr11:67025375-67059633	0.012445	3	*KDM2A*[E]	
8q21.13	60	8q21.13	chr8:80432552-81861219	0.020548	6	*MRPS28*[M]	
2p15	61	2p15	chr2:59143237-63355557	0.021056	25		*XPO1* = 1.1e-05
14q11.2	62	14q11.2[T]	chr14:1-21645085	0.027803	57		
NEDD9	63	6p24.2	chr6:11180426-11620845	0.082606	2	*NEDD9*[K]	
5p13.1	64	5p13.1	chr5:35459650-50133375	0.094657	61		*SLC1A3*=0.0021, *IL7R*=0.0021
LINC00536	65	8q23.3	chr8:116891361-117360815	0.095294	1	*LINC00536*	
10p15.1	66	10p15.1	chr10:4190059-6130004	0.10391	21		
22q11.21	67	22q11.21	chr22:18613558-23816427	0.13213	105		
PHF3	68	6q12	chr6:63883156-64483307	0.17851	4	*PHF3*[E], *EYS*[L]	*PHF3* = 0.051
PAX8	69	2q13	chr2:113990138-114122826	0.19717	2	*PAX8*[K]	
9p24.2	70	9p24.2[T]	chr9:1-7379570	0.20405	45		

Table 2- B) Deletions **Continued**

Peak Name	Rank	Genomic location	Peak region	GISTIC q-value	Gene count	Target(s)	Frequently mutated genes[B]
CDKN2A	1	9p21.3	chr9:21865498-22448737	0	4	*CDKN2A*[K]	*CDKN2A* = 4.4e-15
STK11	2	19p13.3	chr19:1103715-1272039	1.46E-238	7	*STK11*[K]	*STK11* = 2.5e-13
PDE4D	3	5q11.2	chr5:58260298-59787985	2.02E-143	3	*PDE4D*[L]	
PARK2	4	6q26	chr6:161693099-163153207	5.85E-137	1	*PARK2*[L,K]	
LRP1B	5	2q22.1	chr2:139655617-143637838	4.25E-107	1	*LRP1B*[L]	
CSMD1	6	8p23.2	chr8:2079140-6262191	2.39E-96	1	*CSMD1*[L]	
1p36.23	7	1p36.23	chr1:7829287-8925111	1.23E-93	8		
ARID1A	8	1p36.11	chr1:26900639-27155421	5.74E-87	2	*ARID1A*[K]	*ARID1A* = 1.5e-14
PTEN	9	10q23.31	chr10:89615138-90034038	1.12E-79	2	*PTEN*[K]	*PTEN* = 2.2e-15
WWOX	10	16q23.1	chr16:78129058-79627770	8.14E-76	1	*WWOX*[L]	*WWOX* = 0.092
RB1	11	13q14.2	chr13:48833767-49064807	3.88E-75	2	*RB1*[K]	*RB1* = 1.7e-13
FAM190A	12	4q22.1	chr4:90844993-93240505	9.26E-75	1	*FAM190A*[L]	
2q37.3	13	2q37.3[T]	chr2:241544527-243199373	1.77E-70	29	*ING5*[E]	
22q13.32	14	22q13.32[T]	chr22:48026910-51304566	8.20E-65	45	*BRD1*[E], *HDAC10*[E]	
11p15.5	15	11p15.5[T]	chr11:1-709860	1.02E-62	34	*SIRT3*[E], *PHRF1*[E]	*HRAS* = 7.8e-13
LINC00290	16	4q34.3	chr4:178911874-183060693	1.21E-55	1	*LINC00290*	
FHIT	17	3p14.2	chr3:59034763-61547330	3.01E-55	1	*FHIT*[L]	
RBFOX1	18	16p13.3	chr16:5144019-7771745	1.00E-45	1	*RBFOX1*[L]	
PTPRD	19	9p24.1	chr9:8310705-12693402	3.24E-38	1	*PTPRD*[L]	
18q23	20	18q23[T]	chr18:74979706-78077248	1.69E-37	12		
FAT1	21	4q35.2	chr4:187475875-188227950	6.81E-36	1	*FAT1*[K]	*FAT1* = 2.4e-15
MPHOSPH8	22	13q12.11[T]	chr13:1-20535070	2.57E-31	10	*MPHOSPH8*[E]	
15q15.1	23	15q15.1	chr15:41795901-42068054	2.71E-29	4		*MGA* = 0.0083, *RPAP1*=0.035
11q25	24	11q25[T]	chr11:133400280-135006516	4.93E-26	14		
1p13.2	25	1p13.2	chr1:110048528-117687124	1.69E-25	100	*TRIM33*[E]	*NRAS*=1.8e-13, *CD58*=0.079
NF1	26	17q11.2	chr17:29326736-29722618	6.59E-23	5	*NF1*[K]	*NF1* = 3.3e-13
MACROD2	27	20p12.1	chr20:14302876-16036135	9.00E-19	3	*MACROD2*[L]	
7p22.3	28	7p22.3[T]	chr7:1-1496620	1.04E-17	18		
6p25.3	29	6p25.3	chr6:1608837-2252425	3.01E-17	2		
21q11.2	30	21q11.2[T]	chr21:1-15482604	2.34E-14	14		
9p13.1	31	9p13.1	chr9:38619152-71152237	9.75E-14	48		
ZNF132	32	19q13.43[T]	chr19:58661582-59128983	3.77E-13	24	*TRIM28*[E], *ZNF132*	
5q15	33	5q15	chr5:73236070-114508587	8.15E-13	156	*APC*[K], *CHD1*[E]	*APC*=2.6e-13, *RASA1*=0.0029
MLL3	34	7q36.1	chr7:151817415-152136074	9.26E-13	1	*MLL3*[K,E]	*MLL3* = 1.1e-05
19q13.32	35	19q13.32	chr19:47332686-47763284	2.38E-12	10		
15q12	36	15q12[T]	chr15:1-32929863	3.40E-11	155		*OTUD7A* = 0.027
12q24.33	37	12q24.33[T]	chr12:131692956-133851895	1.24E-10	27		*POLE*=3.9e-05, *PGAM5*=0.038
10q26.3	38	10q26.3[T]	chr10:135190263-135534747	2.09E-10	14		
6q21	39	6q21	chr6:86319089-117076132	4.56E-10	141	*PRDM1*[E], *HDAC2*[E], *PRDM13*[E]	*PRDM1* = 0.00054
PPP2R2A	40	8p21.2	chr8:25896447-26250295	1.78E-09	1	*PPP2R2A*	
IKZF2	41	2q34	chr2:211542637-214143899	3.24E-09	4	*IKZF2*[K], ERBB4[L]	*ERBB4* = 0.00058
CNTN4	42	3p26.3[T]	chr3:1-3100786	6.44E-09	3	*CNTN4*[L]	
3p12.2	43	3p12.2	chr3:75363575-86988125	1.22E-07	12	*ROBO1*[L], *CADM2*[L]	
RAD51B	44	14q24.1	chr14:68275375-69288431	1.38E-07	2	*RAD51B*[L]	*ZFP36L1* = 0.0016
11q23.1	45	11q23.1	chr11:105849158-117024891	5.31E-07	84	*ATM*[K]	*ATM*=1.4e-06, *POU2AF1*=0.082
IMMP2L	46	7q31.1	chr7:109599468-111366370	5.74E-07	2	*IMMP2L*[L]	
NEGR1	47	1p31.1	chr1:71699756-74522473	7.25E-07	2	*NEGR1*[L]	
BRCA1	48	17q21.31	chr17:41178765-41336147	7.25E-07	2	*BRCA1*[K]	*BRCA1* = 3.5e-08
9q34.3	49	9q34.3	chr9:135441810-139646221	8.73E-06	94	*NOTCH1*[K], *BRD3*[E], *GTF3C4*[E]	*NOTCH1*=1e-08, *RXRA*=2.1e-05, *COL5A1*=0.0022, *TSC1*=0.012
ANKS1B	50	12q23.1	chr12:99124001-100431272	8.73E-06	2	*ANKS1B*[L]	
DMD	51	Xp21.2	chrX:30865118-34644819	5.15E-05	4	*DMD*[L]	
ZMYND11	52	10p15.3[T]	chr10:1-857150	7.12E-05	4	*ZMYND11*[E]	
PRKG1	53	10q11.23	chr10:52644085-54061437	9.79E-05	3	*PRKG1*[L]	
FOXK2	54	17q25.3	chr17:80443432-80574531	0.00019271	1	*FOXK2*	
AGBL4	55	1p33	chr1:48935280-50514967	0.000219	2	*AGBL4*[L]	
CDKN1B	56	12p13.1	chr12:12710990-12966966	0.00035777	5	*CDKN1B*[K]	*CDKN1B* = 2.2e-06
14q32.33	57	14q32.33[T]	chr14:94381429-107349540	0.00074358	227	*SETD3*[E], *TDRD9*[E]	*AKT1*=2.1e-13, *TRAF3*=9.7e-05
14q11.2	58	14q11.2[T]	chr14:1-30047530	0.0010181	162	*PRMT5*[E], *CHD8*[E]	*CHD8* = 0.034
2p25.3	59	2p25.3[T]	chr2:1-20072169	0.0011137	86	*MYCN*[K]	*MYCN=0.068*
5q35.3	60	5q35.3[T]	chr5:153840473-180915260	0.0028515	212	*NSD1*[E], *ODZ2*[L]	*NPM1*=3.5e-13, *NSD1*=1.9e-09, *ZNF454*=0.0019, *UBLCP1*=0.03, *GABRB2*=0.07
PTTG1IP	61	21q22.3	chr21:46230687-46306160	0.012227	1	*PTTG1IP*	
22q11.1	62	22q11.1[T]	chr22:1-17960585	0.020332	15		
SMAD4	63	18q21.2	chr18:48472083-48920689	0.036866	3	*SMAD4*[K]	*SMAD4* = 6.6e-15
17p13.3	64	17p13.3[T]	chr17:1-1180022	0.040814	16		
4p16.3	65	4p16.3[T]	chr4:1-1243876	0.056345	27		
9p21.2	66	9p21.2	chr9:27572512-28982153	0.091742	3		
10q25.1	67	10q25.1	chr10:99340084-113910615	0.11879	137	*HPSE2*[L], *SMNDC1*[E]	*SMC3*=0.00031, *GSTO2*=0.086
SMYD3	68	1q44	chr1:245282267-247110824	0.15417	8	*SMYD3*[E]	
8p11.21	69	8p11.21	chr8:42883855-47753079	0.17382	4		
Xp22.33	70	Xp22.33[T]	chrX:1-11137490	0.21462	52		*MXRA5* = 0.031

Table 2-Notes Continued

[B]GENE = p-value from [Lawrence et al. *upublished data*] corrected to FDR within peak

[K]Known frequently amplified oncogene or deleted TSG

[P]Putative cancer gene

[E]Epigenetic regulator

[M]Mitochondria-associated gene

**Immediately adjacent to peak region

[T]Adjacent to telomere or centromere of acrocentric chromosome

these regions included 21% of all focal amplifications and 23% of all focal deletions. Focal SCNAs within peak regions tended to be shorter than focal SCNAs elsewhere on the chromosome (median 12.2 Mb in peak regions vs 19.4 Mb genomewide, $p<0.0001$), and were more often high-amplitude events ($p<0.0001$). The number of focal SCNAs involving peak regions per sample tracked the total number of SCNAs ($r=0.84$, $p<0.0001$), ranging from 0.4 focal SCNAs in the typical acute myeloid leukemia to 12.3 focal SCNAs in the typical ovarian cancer (mean 5.2).

Tissue types of similar lineages tended to have similar rates of amplification and deletion in peak SCNA regions (**Fig. 5a**). We observed clusters of squamous cell carcinomas (head and neck squamous cell carcinoma, lung squamous cell carcinoma and bladder cancer) and reproductive cancers (ovarian and endometrial cancer) with breast cancer.

The 70 peak regions of amplification contain a median of three genes each (including microRNAs), with 60 peaks containing fewer than 25 genes. Twenty-four of these peak regions contain an oncogene known to be activated by amplification (**Table 2**), including seven of the top ten regions (*CCND1*, *EGFR, MYC*, *ERBB2, CCNE1*, *MCL1*, and MDM2). The ninth and tenth most significant regions (11q14.1 and 8p11.23, respectively) do not contain known oncogenes, but the latter contains the histone methyltransferase *WHSC1L1* and is 18 kb away from the known amplified oncogene *FGFR1*. The fourth most significantly amplified peak region (3q26.2) contained *TERC*, which encodes the RNA substrate for the known oncogene *TERT*, which is itself in a peak region of amplification (5p15.33). Another peak with eight genes (9p13.3) contain *RMRP*, another *TERT*-associated RNA[14,283].

The 70 peak regions of deletion contain a median of four genes (including microRNAs), with 52 peaks containing fewer than 25 genes. Twenty-two of these regions contain one of the 100 largest genes in the genome and 12 contain known tumor suppressors (**Table 2**; two additional large regions contain the known tumor suppressors *ATM* and *NOTCH1*). Four others each contain a single gene (*PPP2R2A,* *PTTG1IP*, *FOXK2,* and *LINC00290*). We discuss *PPP2R2A* and its binding partner *PPP2R1A* (which is significantly mutated in the same set of cancers[19,26-28 284] in greater detail below. *LINC00290* is a long non-coding RNA, a group whose role in cancer is increasingly being appreciated[29-32,285,286]. Two other regions contain suspected tumor suppressors (*ERRFI1*[28,33-37,287], and *FOXC1*[38-41,288]).

The features most associated with genes in the amplification and deletion peak regions are known to be associated with cancer (**Fig. 5c**). We applied GRAIL[50-52,289], which uses literature citations to find common features of genes in selected regions of the genome. We considered amplifications and deletions separately, and only peaks with fewer than 25 genes.

Among the 37 peak regions of amplification with fewer than 25 genes and without known targets (**Table 2**), the most associated features were related to epigenetic and mitochondrial regulation: “Histone”, “Cytochrome”, “Mitochondrial”, and “Acetyltransferase” (**Fig. 5c**). Thirteen of these 37 regions contain chromatin-state and histone-modifying genes (**Table 2**), reflecting significant enrichment ($p<0.0001$)[40]. Among these, five (*BRD4*, *KAT6A*, *KAT6B*, *NSD1*, and *PHF1*) are subject to recurrent rearrangements in leukemias, sarcomas, and midline carcinomas[26,191,290-295]. The *BRD4* peak also contains *NOTCH3*, another potential oncogene[3,296]. Two others, *KDM2A* and

KDM5A, are reported to regulate the activity of *TP53* and *RB1*, respectively[104,105,297,298]. The finding that multiple peak regions of amplification contain epigenetic regulators is consistent with growing evidence suggesting epigenetic alterations and chromatin remodeling plays a critical role in many forms of cancer[147,299-301]. Ten regions contain genes encoding mitochondria-associated proteins (**Table 2**); none of these are subject to recurrent rearrangements in cancer. The 21 peak regions of deletion with fewer than 25 genes and without known tumor suppressor or large genes were most associated with “Pten”, “Phosphatase”, “Leucine”, and “Prostate”.

Fifty of the 140 peak regions contain a significantly mutated gene, including 23 regions without known oncogene or tumor suppressor gene targets and 32 regions with fewer than 25 genes (**Table 2**). We calculated the significance of mutations (including both point mutations and small insertion-deletion events identified in the paired sequencing data) for each gene in each region using the methods of [19,189,284,302] and corrected for multiple hypotheses reflecting the number of genes in the region. In three cases, there were two significantly mutated genes per peak, for a total of 35 significantly mutated genes. These 35 genes included eight of the 23 known amplification-activated oncogenes and all of the 12 known tumor suppressor genes in these peak regions (**Table 2**). An additional two of the 35 genes (both in amplification peaks) are oncogenes known to be activated by mutations but not by amplifications.

Frame-shift and nonsense mutations that are likely to cause loss of function were significantly enriched in genes in deleted regions (p=0.0002), accounting for 19% of these mutations compared to 12% of mutations found in genes in amplified regions. We excluded regions with known oncogenes or tumor suppressor genes or more than 25

genes from this analysis. These findings are consistent with the prediction that deleted regions without known tumor suppressors are enriched for novel tumor suppressors or genes whose functions are non-essential.

Most peak regions in lineage-specific analyses intersected peak regions in other lineages, and indeed in the Pan-Cancer analysis (**Fig. 6**). We obtained a median of 74 peak regions for each lineage (ranging from 25 in acute myeloid leukemia to 95 in endometrial cancer; 42% were amplification peaks and 58% were deletion peaks), resulting in a total of 770 peak regions. Of these, 84% intersected peak regions in at least one other lineage ($p<0.0001$), and 65% intersected peak regions in the Pan-Cancer analysis. Peak regions tended to be larger in the lineage-specific than the Pan-Cancer analyses (1.4 vs 0.7 Mb), indicating the improved resolution of the Pan-Cancer analysis.

Nevertheless, some significant SCNAs were identified in lineage-specific but not the Pan-Cancer analysis. Across all lineages, we identified 229 peaks not present in the Pan-Cancer analysis, including amplifications of the known amplified oncogenes *MET, CCND2, ERBB3,* and *MYCN* and deletions of the known tumor suppressor genes *TP53* and *CDKN2C*.

Correlations reflect overall levels of genomic disruption

For each pair of peak regions, we looked for positive and negative correlations between focal SCNAs involving these regions (**Fig. 7a**). We compared the number of samples with SCNAs involving both regions between observed data and permuted data in which SCNAs were randomly assigned to samples while maintaining genomic

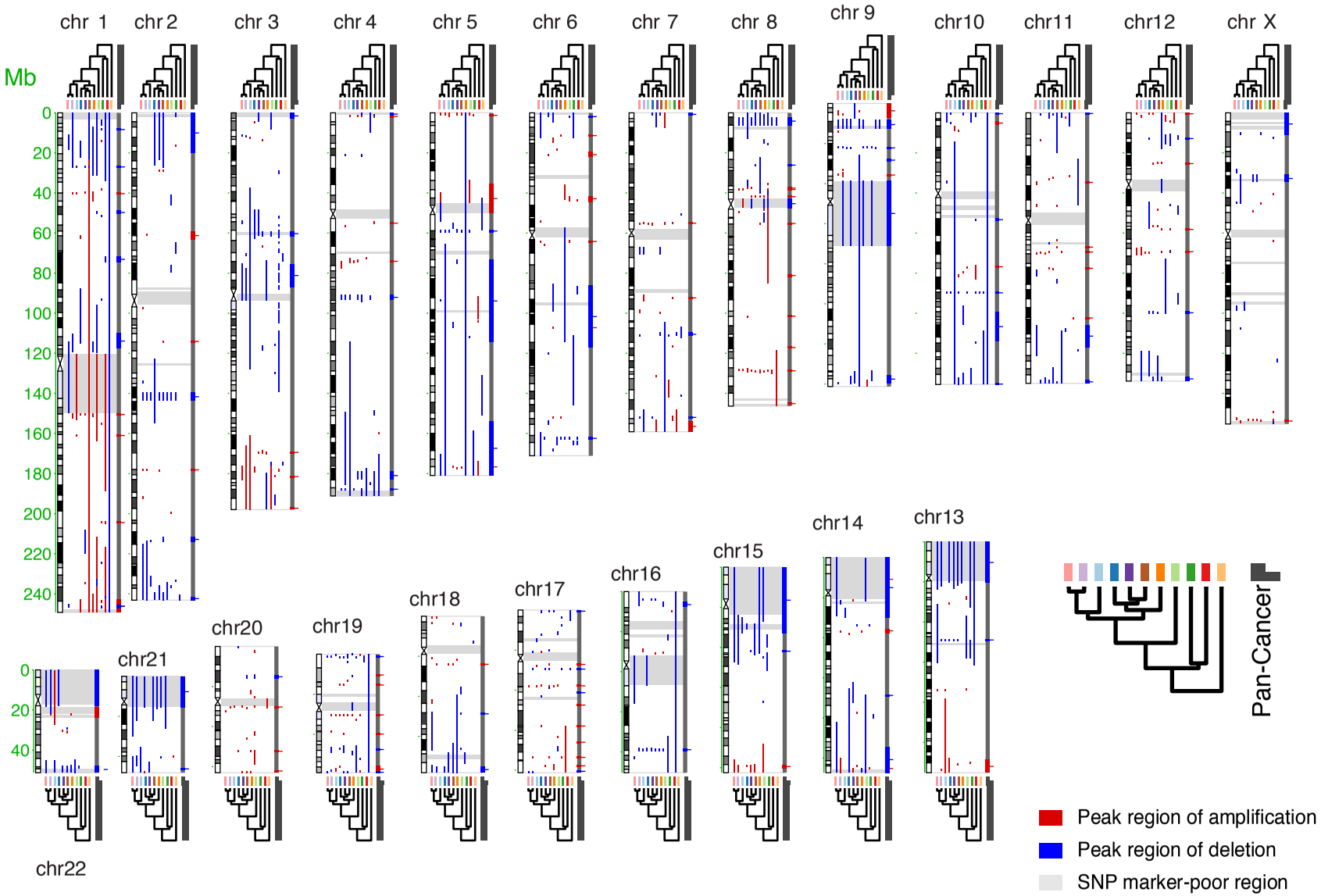


Figure 6- Lineage significance analysis
Illustration of locations of peak regions within all chromosomes (indicated by green numbers) across cancer types (designated by boxes on top and bottom colored according to the scheme on the lower right and reflecting the clustering results in Figure 3a) and the Pancancer analysis (right-most column, denoted by a black line). Detailed information about each peak can be found in Table 3

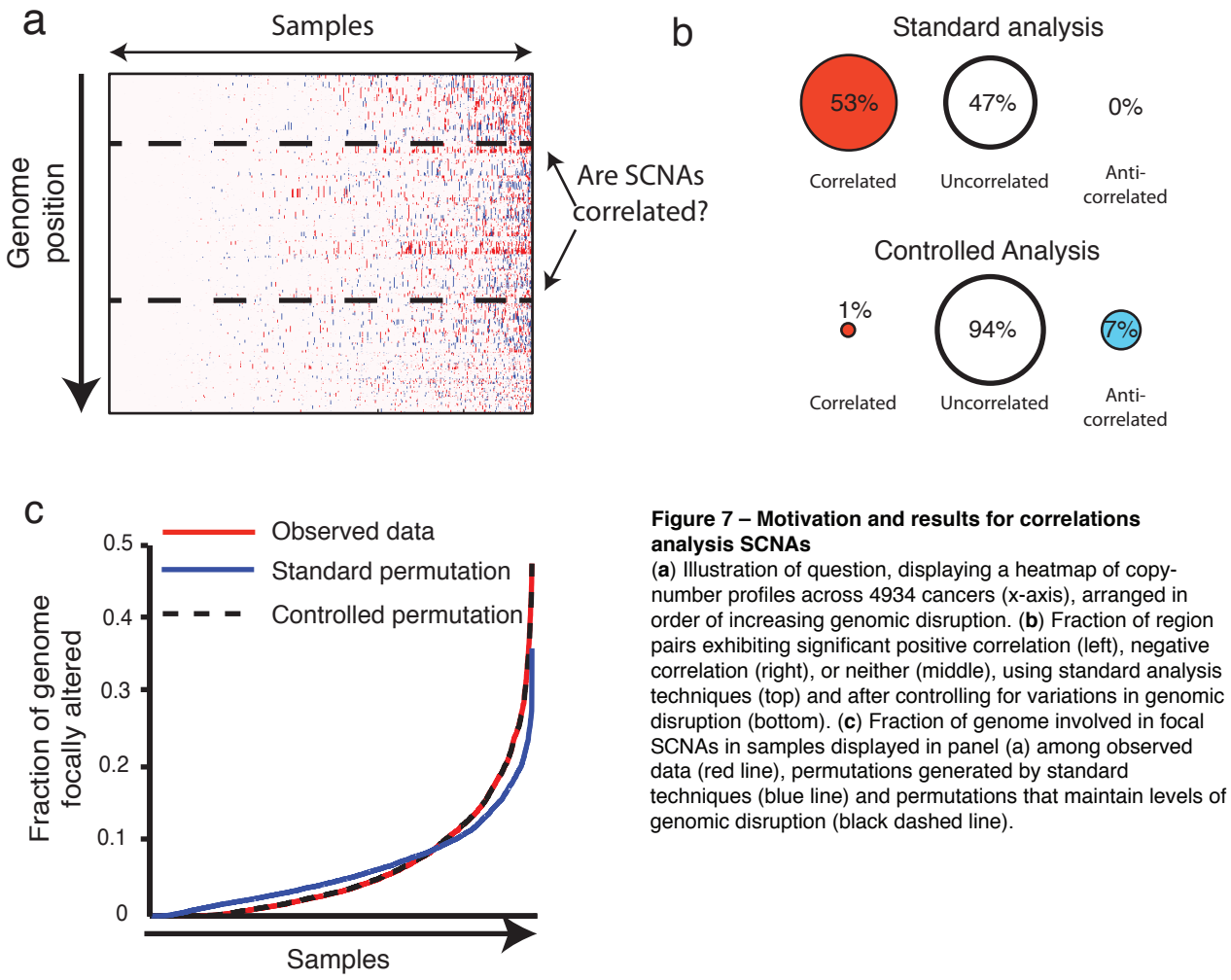


Figure 7 – Motivation and results for correlations analysis SCNAs

(**a**) Illustration of question, displaying a heatmap of copy-number profiles across 4934 cancers (x-axis), arranged in order of increasing genomic disruption. (**b**) Fraction of region pairs exhibiting significant positive correlation (left), negative correlation (right), or neither (middle), using standard analysis techniques (top) and after controlling for variations in genomic disruption (bottom). (**c**) Fraction of genome involved in focal SCNAs in samples displayed in panel (a) among observed data (red line), permutations generated by standard techniques (blue line) and permutations that maintain levels of genomic disruption (black dashed line).

positions and SCNA structure. We only permuted SCNAs within lineages (and sub-lineages when available) to avoid lineage-dependent confounders, and evaluated correlations between regions on different chromosomes to avoid correlations due to chromosomal structure (see Methods). We focused on peak regions with less than 25 genes.

We identified significant positive correlations ($q<0.25$) between 53% of region pairs, but no significant anticorrelations (**Fig. 7b**). The high rate of positive correlations results from widely differing levels of genomic disruption across samples, which are not maintained in permuted datasets (**Fig. 7c**). Similar results are obtained with other standard statistical approaches such as Fisher's exact tests (data not shown). These findings indicate that varying levels of overall genomic disruption confound analyses of functionally relevant correlations between SCNAs.

We therefore re-evaluated correlations between SCNAs after controlling for genomic disruption, by maintaining in the permuted data the fractions of the genome affected by each of amplifications and deletions in each sample (**Fig. 7c**, **Fig. 8a-d**; **Methods**). We performed the analysis in two ways: evaluating all SCNAs, and evaluating only high-level amplifications and homozygous deletions (**Table 3;** see **Methods**). In many cases, high-level amplification or homozygous deletion may be necessary to activate an oncogene or inactivate a tumor suppressor gene[50,137,189,303] and in such cases, correlated features may be masked by noise in lower level events.

When evaluating all SCNAs, we identified significant positive correlations between <1% of region pairs (40 interactions) and anticorrelations between 7% of region pairs (396 interactions, **Fig. 7b**). Correcting for genomic disruption altered the

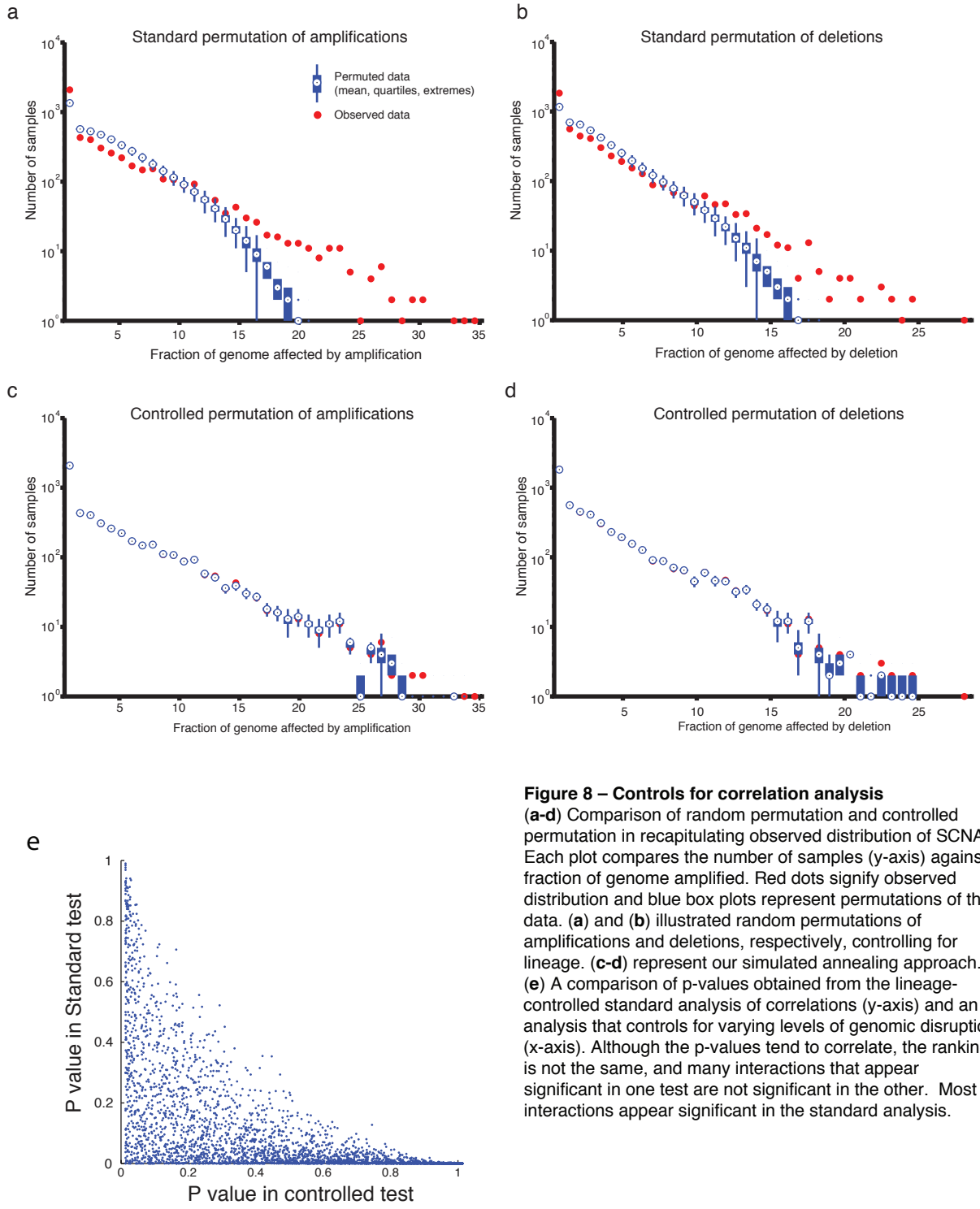


Figure 8 – Controls for correlation analysis
(**a-d**) Comparison of random permutation and controlled permutation in recapitulating observed distribution of SCNAs. Each plot compares the number of samples (y-axis) against fraction of genome amplified. Red dots signify observed distribution and blue box plots represent permutations of this data. (**a**) and (**b**) illustrated random permutations of amplifications and deletions, respectively, controlling for lineage. (**c-d**) represent our simulated annealing approach. (**e**) A comparison of p-values obtained from the lineage-controlled standard analysis of correlations (y-axis) and an analysis that controls for varying levels of genomic disruption (x-axis). Although the p-values tend to correlate, the ranking is not the same, and many interactions that appear significant in one test are not significant in the other. Most interactions appear significant in the standard analysis.

Table 3: high level correlations

p_value	peak1	peak2	q_value	peak1	peak2
0.00012	14	11	0.066707	PDGFRA	CDK4
0.000122	2	71	0.066707	EGFR	CDKN2A
0.000482	57	110	0.175743	19p13.2	PPP2R2A

Table 3: high level anticorrlations Continued

p	peak1	peak2	q	peak1	peak2
0.00012	22	6	0.044472	8q24.3	CCNE1
0.000122	11	71	0.044472	CDK4	CDKN2A
0.00012	14	2	0.044472	PDGFRA	EGFR
0.000241	1	27	0.052723	CCND1	BCL2L1
0.000241	8	71	0.052723	MDM2	CDKN2A
0.000361	21	5	0.056489	IGF1R	ERBB2
0.000361	65	9	0.056489	LINC00536	INTS4
0.000482	1	21	0.058581	CCND1	IGF1R
0.000482	71	81	0.058581	CDKN2A	RB1
0.000964	12	28	0.062027	KAT6A	TUBD1
0.000723	7	29	0.062027	MCL1	[ZNF703]
0.000843	22	32	0.062027	8q24.3	BRD4
0.000843	3	1	0.062027	MYC	CCND1
0.000964	1	6	0.062027	CCND1	CCNE1
0.000964	29	5	0.062027	[ZNF703]	ERBB2
0.000723	2	8	0.062027	EGFR	MDM2
0.000723	110	79	0.062027	PPP2R2A	PTEN
0.001687	29	58	0.090902	[ZNF703]	17q25.1
0.00241	13	58	0.090902	SOX2	17q25.1
0.001566	6	25	0.090902	CCNE1	20q13.33
0.002289	22	1	0.090902	8q24.3	CCND1
0.001928	41	1	0.090902	RMRP	CCND1
0.002289	65	1	0.090902	LINC00536	CCND1
0.002289	2	11	0.090902	EGFR	CDK4
0.002289	30	71	0.090902	1q23.3	CDKN2A
0.001807	4	71	0.090902	TERC	CDKN2A
0.00241	19	5	0.090902	KDM5A	ERBB2
0.002169	15	8	0.090902	BDH1	MDM2
0.002169	76	79	0.090902	CSMD1	PTEN
0.002771	1	58	0.101052	CCND1	17q25.1
0.002892	3	118	0.102044	MYC	BRCA1
0.003012	21	6	0.102974	IGF1R	CCNE1
0.003373	12	54	0.108547	KAT6A	17q21.33
0.003373	30	2	0.108547	1q23.3	EGFR
0.003614	6	36	0.112978	CCNE1	ZNF217
0.003976	22	8	0.117558	8q24.3	MDM2
0.003976	4	8	0.117558	TERC	MDM2
0.004096	38	6	0.117933	KAT6B	CCNE1
0.004458	63	1	0.118948	NEDD9	CCND1

Table 3: high level anticorrlations Continued

p	peak1	peak2	q	peak1	peak2
0.004458	28	6	0.118948	TUBD1	CCNE1
0.004337	55	6	0.118948	BPTF	CCNE1
0.004699	10	19	0.119546	WHSC1L1	KDM5A
0.004699	3	88	0.119546	MYC	RBFOX1
0.00506	59	6	0.12302	KDM2A	CCNE1
0.00506	7	10	0.12302	MCL1	WHSC1L1
0.006506	10	47	0.15473	WHSC1L1	CBX8
0.007229	30	29	0.158169	1q23.3	[ZNF703]
0.007108	10	55	0.158169	WHSC1L1	BPTF
0.006988	9	21	0.158169	INTS4	IGF1R
0.006988	17	89	0.158169	MDM4	PTPRD
0.007831	8	6	0.164759	MDM2	CCNE1
0.007831	63	9	0.164759	NEDD9	INTS4
0.008554	3	136	0.166493	MYC	9p21.2
0.008313	22	9	0.166493	8q24.3	INTS4
0.008072	13	8	0.166493	SOX2	MDM2
0.008675	17	10	0.166493	MDM4	WHSC1L1
0.008434	3	36	0.166493	MYC	ZNF217
0.009277	4	11	0.17202	TERC	CDK4
0.009157	4	80	0.17202	TERC	WWOX
0.009759	29	54	0.175023	[ZNF703]	17q21.33
0.009759	10	28	0.175023	WHSC1L1	TUBD1
0.010241	30	1	0.177835	1q23.3	CCND1
0.010241	65	33	0.177835	LINC00536	KRAS
0.010723	7	1	0.180475	MCL1	CCND1
0.010723	13	3	0.180475	SOX2	MYC
0.011084	11	136	0.183731	CDK4	9p21.2
0.012169	10	58	0.185243	WHSC1L1	17q25.1
0.012169	20	22	0.185243	MYCL1	8q24.3
0.012289	12	55	0.185243	KAT6A	BPTF
0.011928	15	71	0.185243	BDH1	CDKN2A
0.011687	35	71	0.185243	NFE2L2	CDKN2A
0.012048	1	91	0.185243	CCND1	FAT1
0.01253	32	23	0.185243	BRD4	G6PD
0.01241	3	9	0.185243	MYC	INTS4
0.012892	14	6	0.185571	PDGFRA	CCNE1
0.012771	13	36	0.185571	SOX2	ZNF217
0.013253	13	92	0.185882	SOX2	MPHOSPH8
0.013133	34	71	0.185882	NKX2-1	CDKN2A

Table 3: high level anticorrlations Continued

p	peak1	peak2	q	peak1	peak2
0.013494	15	9	0.186866	BDH1	INTS4
0.014096	74	71	0.190388	PARK2	CDKN2A
0.014096	9	36	0.190388	INTS4	ZNF217
0.014819	20	15	0.192283	MYCL1	BDH1
0.01494	65	32	0.192283	LINC00536	BRD4
0.014819	20	32	0.192283	MYCL1	BRD4
0.014699	3	124	0.192283	MYC	FOXK2
0.016867	71	114	0.21457	CDKN2A	RAD51B
0.017711	56	1	0.217704	E2F3	CCND1
0.017711	3	11	0.217704	MYC	CDK4
0.017349	49	10	0.217704	CDK6	WHSC1L1
0.018916	71	137	0.22539	CDKN2A	10q25.1
0.020602	22	54	0.22539	8q24.3	17q21.33
0.020361	3	54	0.22539	MYC	17q21.33
0.019518	9	28	0.22539	INTS4	TUBD1
0.018795	15	55	0.22539	BDH1	BPTF
0.02	86	71	0.22539	LINC00290	CDKN2A
0.020482	12	21	0.22539	KAT6A	IGF1R
0.019759	10	21	0.22539	WHSC1L1	IGF1R
0.019759	13	19	0.22539	SOX2	KDM5A
0.02012	71	79	0.22539	CDKN2A	PTEN
0.019277	4	88	0.22539	TERC	RBFOX1
0.021084	69	13	0.22614	PAX8	SOX2
0.021084	30	10	0.22614	1q23.3	WHSC1L1
0.021325	12	36	0.226504	KAT6A	ZNF217
0.021687	3	8	0.228128	MYC	MDM2
0.023373	3	59	0.24353	MYC	KDM2A
0.024217	71	123	0.246528	CDKN2A	PRKG1
0.024337	13	71	0.246528	SOX2	CDKN2A
0.024217	4	72	0.246528	TERC	STK11

estimated significance of these interactions and also changed the rank ordering of those significance estimates (**Fig. 8e**). High-level amplifications and homozygous deletions are relatively rare, limiting our power to detect anticorrelations in the high-level analysis. Among the 1094 interactions we were powered to detect, we observed positive correlations between <1% of region pairs (3 interactions, **Table 3**) and anticorrelations between 10% of region pairs (108 interactions, **Fig. 9a**, **Table 3**). The three correlations included deletions of *CDKN2A* with amplifications of *EGFR*, amplifications of *PDGFR* with amplifications of *CDK4*, and deletions of *PPP2RA* with amplifications of 19p13.2.

We predicted that anticorrelated SCNAs would often indicate functional redundancies, and therefore genes in the affected regions would often be in similar pathways and interact physically. We tested this hypothesis by comparing networks representing significantly anticorrelated SCNAs ("anticorrelation networks") with DAPPLE, a set of curated protein-protein interactions (PPIs)[50,191,192,279] (see **Methods**).

Networks formed by our anticorrelations analyses and by PPIs significantly overlapped ($p<0.0001$ and $p=0.006$ for all-SCNA and high-level analyses, respectively, **Fig. 9b-c**). For example, in the analysis of all SCNAs, we observed 100 overlapping edges, a 2-fold increase over the 43.4 overlapping edges expected by chance. This significance was not observed for correlated events ($p=1$ for both all-SCNA and high–level analyses). These results suggest that the observed anticorrelations are related to biological interactions.

The anticorrelations networks were enriched for both isolated nodes and highly connected "hub" regions (**Fig. 9d**). To analyze the structure of these networks, we generated control anticorrelation networks representing the most significant edges from

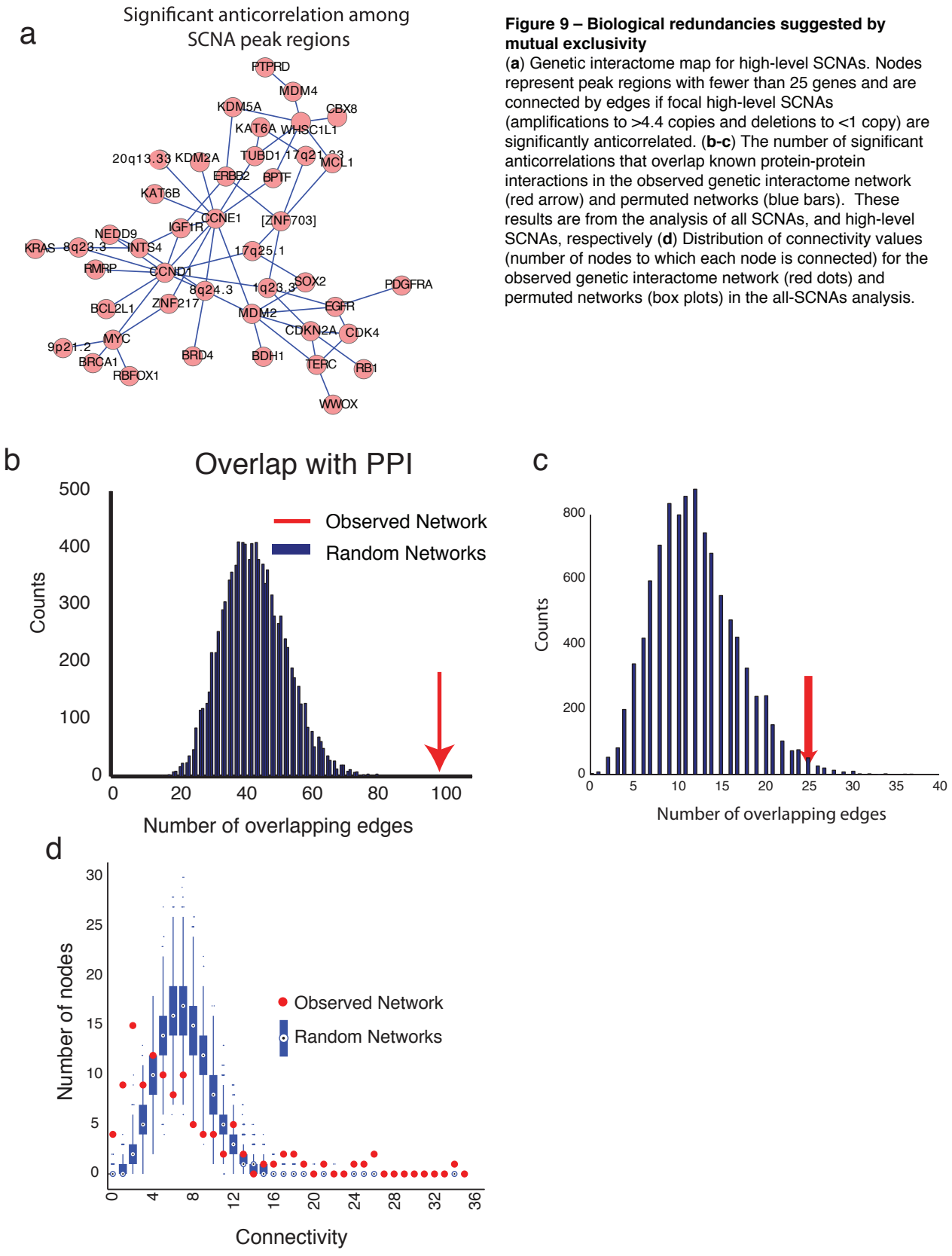


Figure 9 – Biological redundancies suggested by mutual exclusivity
(**a**) Genetic interactome map for high-level SCNAs. Nodes represent peak regions with fewer than 25 genes and are connected by edges if focal high-level SCNAs (amplifications to >4.4 copies and deletions to <1 copy) are significantly anticorrelated. (**b-c**) The number of significant anticorrelations that overlap known protein-protein interactions in the observed genetic interactome network (red arrow) and permuted networks (blue bars). These results are from the analysis of all SCNAs, and high-level SCNAs, respectively (**d**) Distribution of connectivity values (number of nodes to which each node is connected) for the observed genetic interactome network (red dots) and permuted networks (box plots) in the all-SCNAs analysis.

permuted data in which we had randomized the SCNA sample assignments within lineage. In the all-SCNA analysis, 28 regions were anticorrelated with fewer than three other regions, relative to three isolated nodes in the average permutation ($p<0.01$).

The isolated nodes in the all-SCNA analysis were enriched for regions containing large genes (including 10 of 28 such regions; $p=0.004$). Conversely, they trended toward excluding regions with known oncogenes or tumor suppressors (five of 35 such regions; $p=0.06$). Most peak regions exhibit fewer anticorrelations in the high-level analysis, possibly due to decreased power. The most extreme exception was CDKN2A, which anticorrelated with 14 regions in the high-level analysis and only nine regions in the all-SCNA analysis. Consistent with these findings, *CDKN2A* is often inactivated by homozygous deletions.

Whole genome doubling: correlates and consequences on cancer evolution

We applied a similar analysis to identify events associated with WGD. We included both SCNAs and mutations, using the 200 most significantly mutated genes across the TCGA Pan-Cancer dataset[284,302]. Three SCNA peak regions and two significantly mutated genes correlated with WGD (**Table 4**). *TP53* mutations and CCNE1 amplifications correlated with WGD; both have been functionally associated with tolerance of tetraploidy in experimental models[304-306]. Our findings indicate these associations apply to human tumors across multiple lineages. We also found that deletions of *PPP2R2A* and mutations of its binding partner *PPP2R1A* were correlated with WGD. These two genes belong to phospho-protein phosphatase complex 2 (PPP2), which regulates mitotic spindle formation and can lead to chromosomal missegregation and abnormal mitoses when depleted[307,308].

Table 4: Correlates with WGD

putative gene	Rank	p_value
Anticorrelations with WGD		
Amplications		
MYC	3	0.0065
8q24.3	22	0.0007
Deletions		
CDKN2A	1	0
PTEN	9	0
10q25.1	67	0.0016
NF1	26	0.0053
DMD	51	0
Mutations		
PIK3CA	2	0.002
CTCF	28	0.004
MAP3K1	45	0.004
ATM	68	0.006

putative gene	Rank	p_value
Correlations with WGD		
Amplications		
CCNE1	6	0.00012
Deletions		
CSMD1	6	0.00012
PPP2R2A	40	0.00012
Mutations		
TP53	1	<0.001
PPP2R1A	84	0.008

Eleven genetic events anti-correlated with WGD, including two amplifications, five deletions and four mutations (**Table 4**). The deletions included *CDKN2A*, *PTEN*, and *NF1*, and three of the four mutations also involved genes known as or proposed to be tumor suppressors (*CTCF*[309], *MAP3K1*[310], and *ATM*). The anticorrelations of these tumor suppressors may result from a greater difficulty in biallelically inactivating tumor suppressors in samples with extra copies subsequent to WGD[147].

We can use the unique signature that WGD places on the cancer genome to determine when specific regions were lost during cancer development. A loss of heterozygosity (LOH) deletion in WGD samples will appear to be a single event of amplitude two (going from two copies to zero copies). We hypothesized that TSGs important to tumor initiation are more likely to occur prior to WGD. Conversely, the timing of LOH in peaks whose significance may be derived from increased SCNA generation would be similar to timing of events observed throughout the rest of the genome. We looked at what fraction of focal LOH on each significant peak occurred prior to WGD and compared these numbers to overall fraction of LOH observed prior to WGD.

We found that LOH in well-studied TSGs occurred relatively early in cancer development, whereas many significant regions thought to be caused by fragile sites or lack of negative selection were not significantly different from background (**Table 5**). These groups were evenly distributed across significant region list when sorted by GISTIC significance (Kolmogorov-Smirnov test, p=0.27 and 0.86 for TSGs and large genes, respectively). However, regions that were more likely to occur prior to WGD were significantly enriched for known TSGs (one-sided KS test, p = 0.007), with the top

Table 5: LOH timing relative to WGD

Peak Name	GISTIC Rank	Genomic location	GISTIC q-value	Gene count	Target(s)	Frequently mutated genes[B]	wgd timing q value
CDKN2A	1	9p21.3	0	4	CDKN2A[K]	*CDKN2A* = 4.4e-15	2.85E-15
CSMD1	6	8p23.2	2.39E-96	1	CSMD1[L]		9.23E-14
PPP2R2A	40	8p21.2	1.78E-09	1	PPP2R2A		6.55E-09
FOXK2	54	17q25.3	0.00019271	1	FOXK2		1.08E-05
BRCA1	48	17q21.31	0.000000725	2	BRCA1[K]	*BRCA1* = 3.5e-08	0.0002
15q15.1	23	15q15.1	2.71E-29	4		*MGA* = 0.0083, *RPAP1*=0.035	0.0005
9p21.2	66	9p21.2	0.091742	3			0.0014
RB1	11	13q14.2	3.88E-75	2	RB1[K]	*RB1* = 1.7e-13	0.0059
NF1	26	17q11.2	6.59E-23	5	NF1[K]	*NF1* = 3.3e-13	0.0059
PTEN	9	10q23.31	1.12E-79	2	PTEN [K]	*PTEN* = 2.2e-15	0.0174
7p22.3	28	7p22.3[T]	1.04E-17	18			0.0196
21q11.2	30	21q11.2[T]	2.34E-14	14			0.0196
RAD51B	44	14q24.1	0.000000138	2	RAD51B[L]	*ZFP36L1* = 0.0016	0.0196
19q13.32	35	19q13.32	2.38E-12	10			0.0248
STK11	2	19p13.3	1.46E-238	7	STK11[K]	*STK11* = 2.5e-13	0.0276
PTPRD	19	9p24.1	3.24E-38	1	PTPRD[L]		0.032
SMAD4	63	18q21.2	0.036866	3	SMAD4[K]	*SMAD4* = 6.6e-15	0.0462
CDKN1B	56	12p13.1	0.00035777	5	CDKN1B[K]	*CDKN1B* = 2.2e-06	0.0468
IMMP2L	46	7q31.1	0.000000574	2	IMMP2L[L]		0.1279
MACROD2	27	20p12.1	9E-19	3	MACROD2[L]		0.1378
18q23	20	18q23[T]	1.69E-37	12			0.1431
MLL3	34	7q36.1	9.26E-13	1	MLL3[K,E]	*MLL3* = 1.1e-05	0.1431
ANKS1B	50	12q23.1	0.00000873	2	ANKS1B[L]		0.1431
FHIT	17	3p14.2	3.01E-55	1	FHIT[L]		0.1721
10q26.3	38	10q26.3[T]	2.09E-10	14			0.1721
3p12.2	43	3p12.2	0.000000122	12	)BO1[L], *CADM2*[L]		0.1721
11q25	24	11q25[T]	4.93E-26	14			0.179
ZNF132	32	19q13.43[T]	3.77E-13	24	IIM28[E], *ZNF132*		0.179
PTTG1IP	61	21q22.3	0.012227	1	PTTG1IP		0.179
IKZF2	41	2q34	3.24E-09	4	ZF2[K], ERBB	*ERBB4* = 0.00058	0.2653
RBFOX1	18	16p13.3	1E-45	1	RBFOX1[L]		0.3041
ZMYND11	52	10p15.3[T]	0.0000712	4	ZMYND11[E]		0.3402
17p13.3	64	17p13.3[T]	0.040814	16			0.3431
PARK2	4	6q26	5.85E-137	1	PARK2[L,K]		0.4359
PDE4D	3	5q11.2	2.02E-143	3	PDE4D[L]		0.4561
22q11.1	62	22q11.1[T]	0.020332	15			0.7488
WWOX	10	16q23.1	8.14E-76	1	WWOX[L]	*WWOX* = 0.092	0.8334
SMYD3	68	1q44	0.15417	8	SMYD3[E]		0.8334
1p36.23	7	1p36.23	1.23E-93	8			0.8611
FAT1	21	4q35.2	6.81E-36	1	FAT1[K]	*FAT1* = 2.4e-15	0.8869
CNTN4	42	3p26.3[T]	6.44E-09	3	CNTN4[L]		0.8869
8p11.21	69	8p11.21	0.17382	4			0.8869
NEGR1	47	1p31.1	0.000000725	2	NEGR1[L]		0.9142
6p25.3	29	6p25.3	3.01E-17	2			0.9275
ARID1A	8	1p36.11	5.74E-87	2	ARID1A[K]	*ARID1A* = 1.5e-14	0.9326
FAM190A	12	4q22.1	9.26E-75	1	FAM190A[L]		0.9326
MPHOSPH8	22	13q12.11[T]	2.57E-31	10	MPHOSPH8[E]		0.9326
LRP1B	5	2q22.1	4.25E-107	1	LRP1B[L]		0.9604
LINC00290	16	4q34.3	1.21E-55	1	LINC00290		0.9604
PRKG1	53	10q11.23	0.0000979	3	PRKG1[L]		0.9604
AGBL4	55	1p33	0.000219	2	AGBL4[L]		0.9604

10 most significant regions including *RB1*, *CDKN2A*, *NF1*, *PTEN*, and *BRCA1* (**Table 5**). This enrichment was not observed for significant regions containing large genes (p=0.94). PPP2R2A, whose deletion was found to correlate with WGD above, was the second most significant region with respect to early LOH (FDR p-value $9x10^{-14}$). The sixth most significant peak in this analysis (FDR p-value = $5x10^{-4}$) contains 4 genes, including Max-gene associated protein (MGA), which is also significantly mutated across the same dataset.

Discussion

Significant regions of SCNA in the Pancancer dataset

Using current significance techniques, correcting for sample purity and ploidy, we found 140 significant regions in the Pan-Cancer analysis, only 35 of which contained known amplified oncogenes or tumor suppressor genes. As mentioned above, some of the remaining regions may be false positive SCNAs and some of the remaining regions may recur because these regions are subject to relatively small amounts of negative selection or due to mechanistic biases favoring the generation of SCNAs in these regions. Indeed, we found that SCNAs involving large genes (potentially less negative selection) or significant regions near telomeres (mechanistic bias) often did not anticorrelate with any other genetic events, suggesting the genes in these regions may have limited functional roles in oncogenesis. However, it remains likely that many additional oncogenes and tumor suppressor genes are within these regions. Moreover, these 140 regions and the additional 229 peak regions identified in the lineage-specific analyses are likely to compose a subset of the regions that are significantly altered in

cancer. Analyses of other cancer types have identified additional peak regions, and the limited resolution of the array platform may have obscured detection of some SCNAs.

Varying levels of genomic disruption across cancers are likely to engender biases in analyses of correlations not only between SCNAs, but also between SCNAs and other features of these cancers. For example, increased genomic disruption has been associated with poor prognosis in multiple cancer types[253]. Poor prognosis is therefore likely to be associated with increased rates of SCNA across much of the genome. Controlling for this tendency is required to correctly identify SCNAs that are functionally associated with progression. It will also be important to account for other possible confounders, such as mechanistically linked events (e.g. chromothripsis or SCNAs that encompass multiple peak regions). Unlike other analysis, our method controls for the background rate of SCNAs in each sample and does not depend on previous determination of significant regions or genes, which should increase the stability of our results[202,261].

WGD frequency and timing.

Besides *TP53,* we found four genetic events that correlated with WGD. The exact role of WGD in cancer progression, as either a driver event or a frequent consequence of an unstable genome, is still poorly understood, but these correlates may provide clues as to potential causes of WGD. The PPP complex role in tumor suppression has been suggested previously, and having two components in the 5 genes the correlate with WGD suggests its role in guiding and stabilizing mitosis may explain how loss of function can lead to tumor formation.

While SCNAs on many tumor suppressors tended to anti-correlate with WGD, using the timing of SCNAs in relation to the WGD event could be a useful tool in identifying significant regions likely involved early in tumor development. Using the WGD to temporally sequence events in this manner could be incorporated in specific significance analyses, with earlier events providing more evidence of a “tumor initiator” role for a given alteration. We found that events on well-studied TSG genes frequently occurred early in cancer development compared to other genes, whereas many regions whose significance may be due to mechanism where not temporally biased. This analysis identified a couple peaks that exhibited patterns more closely associated with known tumor suppressors, suggesting a potential role in tumor formation.

Chapter 4: The discovery of non-driver cancer dependencies using high-throughput shRNA pooled screens.

Section goals

While finding specific vulnerabilities of primary tumors through identifying the drivers of cancer evolution has been remarkably successful, comprehensive, personalized medicine may require a more inclusive approach to identifying therapeutic targets. Most cancer malignancies are thought to be the result of a small handful of driver alterations that together overcome the many barriers to unrestrained growth. Focusing therapies only on driver alterations may mean that, for each patient, we will have a small handful of opportunities to find specific vulnerabilities that are a direct response of these driver alterations. Many driver alterations, such as amplifications of transcription factor or inactivation of tumor suppressor genes, have proven refractory to current techniques of small molecule inhibition, raising the possibility that some cancers will not have driver alterations that expose tractable therapeutic targets. Expanding the universe of potential therapeutic targets is therefore a worthwhile endeavor. The goal of this section was to discover vulnerabilities predicted by passenger alterations by integrating an RNAi pooled screen across many cell lines with genome-wide SNP6.0 array SCNA data to determine if partial copy-number loss of specific genes renders cells highly dependent on the remaining copy. We identified a class of genes that render cells that harbor copy-number loss highly dependent on the expression of the remaining copy. These were enriched for cell essential genes, most predominantly components of the proteasome, spliceosome, and ribosome.

Methods

Copy-number and High Methylation analysis of Tumor Samples

Copy-numbers were determined for 3,131 cancer samples using Affymetrix 250K SNP array data as previously described[311]. Marker and gene locations were based on the hg18 genome build. We considered markers with relative $\log_2$ copy number ratios less than -0.1 to be affected by partial copy number loss, and markers less than -1.28 as homozygous deletions[26]. Copy-number profiles, and the locations, lengths, and amplitudes of the amplification and deletion events underlying these profiles, were determined as previously described[311]. We determined the significance of depletion of homozygous deletions among candidate CYCLOPS genes by comparing observed rates of homozygous deletion to the distribution of rates after permutation of gene names. DNA methylation state Beta-values were collected for 601 ovarian tumors from the TCGA web portal. Gene level Beta values > 0.7 were considered high DNA methylation.

Copy number analysis of cancer cell lines from CCLE

For each sample, we created a 100 bin histogram of copy number values for all markers, and then used a 5-bin moving average to smooth this distribution. This procedure typically yielded 2-5 well-separated peaks (local maxima with height as measured from local max to surrounding local minima >2% of genome), presumably corresponding to integer level copy loss and gains. Based on these peaks, samples were separated into one of two categories for classification. If a sample contained one peak between $\log_2$ copy number -0.05 and 0.05, with a second peak between -0.05 and

-0.4, the first peak was defined as copy neutral, and the second peak as partial copy loss. In this case, the cutoff for copy loss was set at 95% upper bound of the second peak, and the cutoff for copy neutral was set at 95% lower bound of the first peak. If there were no peak that met our height criteria within these regions in a given sample, markers <-0.4 were considered copy loss while markers >-0.2 were considered copy neutral. In either case, markers that lay between our two cutoffs were left uncalled and genes with these copy numbers were excluded from further analysis. Markers with $\log_2$ copy number ratios ≤ -1.28 were considered homozygous loss and genes with these copy numbers were also removed from further analyses. We used the Kolmogorov-Smirnov test to determine enrichment of the CYCLOPs genes identified in our original analysis among the most significant genes in our analysis of the Validation dataset.

Analysis of copy number and expression correlations

Quantized normalized expression data was obtained from the CCLE (www.broadinstitute.org/ccle) and TCGA (https://tcga-data.nci.nih.gov/tcga/) portals. Enrichment of Pearson correlation coefficients among CYCLOPS candidates and pathways was determined by permuting gene names. A similar analysis was used to determine significance of correlation between Bortezomib logIC50 data for 133 cancer cell lines collected from Sanger center Cancer Genome Project (http://www.sanger.ac.uk/genetics/CGP/) portal and the expression patterns for these lines from the CCLE (www.broadinstitute.org/ccle).

CYCLOPS analysis

For each cell line, we classified each gene as intact (no copy-number loss), partial loss, or to be excluded (for genes undergoing homozygous loss or with ambiguous data) based on thresholds determined using the distribution of relative copy-numbers generated from analysis of SNP array data for that cell line. Gene dependency scores were determined using the ATARiS algorithm[253]. The statistical significance of the difference in mean gene dependency scores between "intact" and "partial loss" cell lines was determined by comparing the observed data to data representing 50,000 random permutations of class labels, each maintaining the number of cell lines and lineage distribution in each class. Multiple hypotheses were corrected using the FDR framework.

Author contributions:

The author was responsible for the data curation, quality control, and analysis, with advice from Deepak Nijhawan, William Hahn and Rameen Beroukhim.

Results

By analyzing copy-number profiles from 3,131 cancers across a wide diversity of cancer types[26], we found that most cancers exhibit relative copy-number loss affecting at least 11% of the genome and that many cancers exhibit much more extensive loss of genetic material (**Fig. 10a**). Here we note that the phenotypic effects of SCNA may be dosage dependent, such that a single gain or loss may represent a larger deviation from "balance" in a diploid cell (2n) than cells with higher ploidy. With this in mind, we used data with DNA intensity normalized to the total DNA content of the sample. Much of this widespread genomic disruption is due to copy-number alterations involving whole chromosomes or chromosome arms, presumably due to mechanisms that favor the

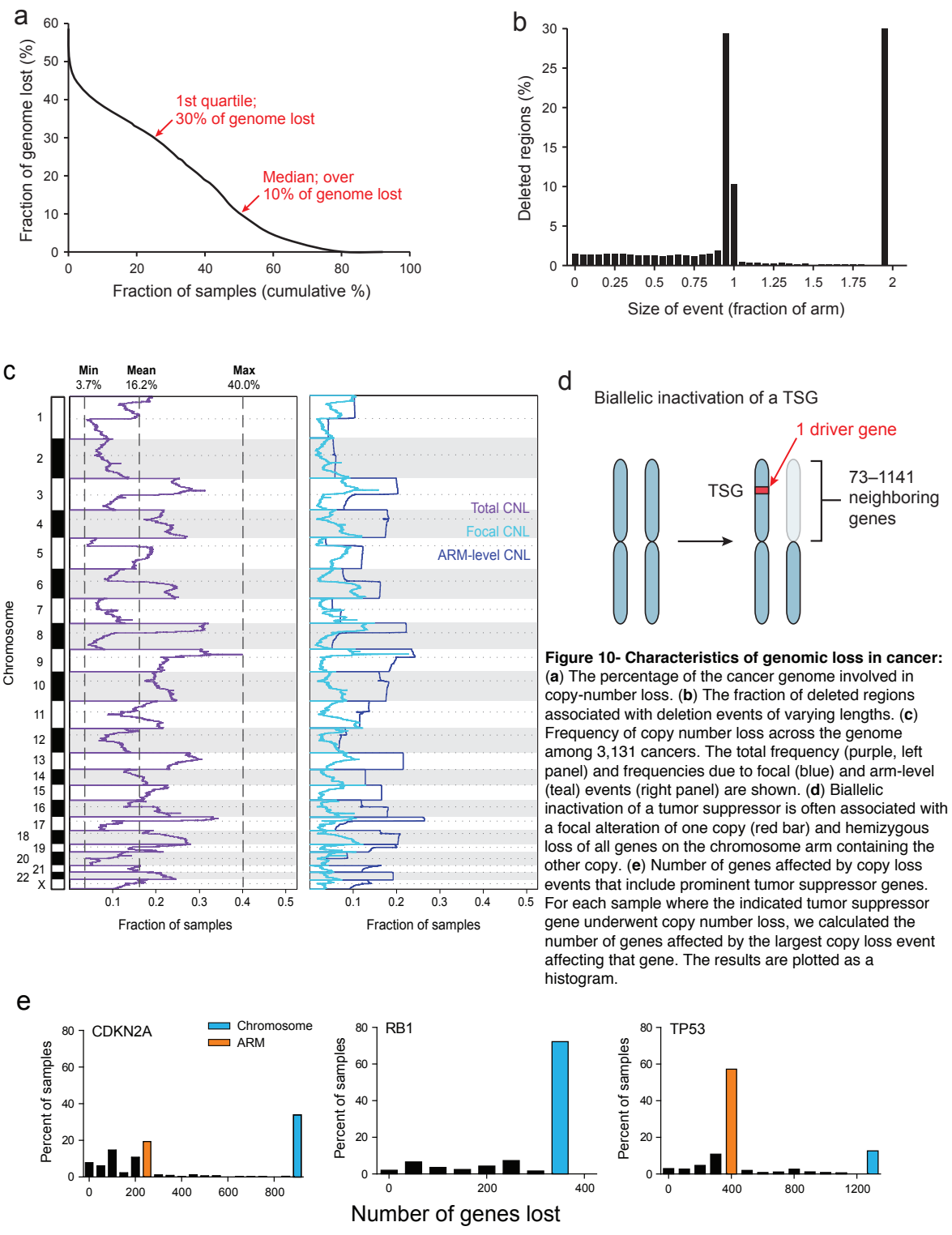


Figure 10- Characteristics of genomic loss in cancer: (**a**) The percentage of the cancer genome involved in copy-number loss. (**b**) The fraction of deleted regions associated with deletion events of varying lengths. (**c**) Frequency of copy number loss across the genome among 3,131 cancers. The total frequency (purple, left panel) and frequencies due to focal (blue) and arm-level (teal) events (right panel) are shown. (**d**) Biallelic inactivation of a tumor suppressor is often associated with a focal alteration of one copy (red bar) and hemizygous loss of all genes on the chromosome arm containing the other copy. (**e**) Number of genes affected by copy loss events that include prominent tumor suppressor genes. For each sample where the indicated tumor suppressor gene underwent copy number loss, we calculated the number of genes affected by the largest copy loss event affecting that gene. The results are plotted as a histogram.

generation of such large events (**Fig. 10b**). As a consequence, most genes undergo copy-number loss in a substantial fraction of cancers (average 16.2, range 3.7-40.2%; **Fig. 10c**). A subset of the genes affected by recurrent copy-number alterations contribute to cancer development as tumor suppressor genes; however, a substantial fraction of these genes are recurrently lost due to passenger events or because of their proximity to a frequently deleted tumor suppressor gene (**Fig. 10d-e**). We hypothesized that for a subset of non-driver genes, hemizygous loss may be tolerated and frequent but complete loss would lead to cell death. In some of these cases, hemizygous loss might lead to sensitivity to further inhibition of the gene relative to cells that harbor two copies of these genes.

To identify genes whose loss correlated with a greater sensitivity to further gene suppression, we integrated gene dependencies and copy-number data from 86 cancer cell lines (**Table 6**) in Project Achilles, described above. Out of the 111 lines available in the Achilles dataset, 101 of the barcode arrays passed internal quality control, and 86 of those also contained Affymetrix SNP 6.0 array data from the Cancer Cell Line Encyclopedia (CCLE) [233,312]. For 7,250 of these genes, the Ataris algorithm[253] determined there were multiple shRNAs that had comparable effects across cell lines, suggesting their effects were due to suppression of the intended target and used these shRNAs to construct composite “gene dependency scores”. For each gene, we first classified each cell line by whether or not it had copy-number loss in that gene and then calculated the mean gene dependency score among cell lines in each class. We then determined the difference in mean scores between the copy-loss and copy-neutral classes and rated the significance of this difference by permuting class labels (**Fig. 11**).

Table 6-Cell Lines used in Achilles CYCLOPS analysis

Cell line	Tissue of Origin	Cell line	Tissue of Origin
786O	kidney	HEC1A	endometrium
A2780	ovary	COLO741	skin
OVCAR4	ovary	KYSE510	esophagus
L33	pancreas	LS411N	large intestine
HS944T	skin	KYSE150	esophagus
CAOV3	ovary	KYSE450	esophagus
HT29	large intestine	LOVO	large intestine
NCIH1975	lung	KYSE30	esophagus
A549	lung	AGS	stomach
NCIH196	lung	U251MG	cns
PANC0327	pancreas	MIAPACA2	pancreas
IGROV1	ovary	NCIH2122	lung
OVCAR8	ovary	KMS12BM	leukemia
TOV21G	ovary	OV90	ovary
KM12	large intestine	NCIH2171	lung
OE33	esophagus	DLD1	large intestine
MDAMB453	breast	HL60	leukemia
HLF	liver	HEYA8	ovary
SJSA1	bone	TYKNU	ovary
LS513	large intestine	EFO21	ovary
HCC827	lung	RMGI	ovary
COV362	ovary	BXPC3	pancreas
NCIH661	lung	EFO27	ovary
CFPAC1	pancreas	ASPC1	pancreas
HPAC	pancreas	RT112	urinary tract
SKCO1	large intestine	A2058	skin
NCIH82	lung	KURAMOCHI	ovary
OVISE	ovary	SW480	large intestine
OVMANA	ovary	IGR39	skin
NCIH1650	lung	CAOV4	ovary
NIHOVCAR3	ovary	COLO205	large intestine
KP4	pancreas	C2BBE1	large intestine
HUG1N	stomach	SNUC1	large intestine
TE9	esophagus	SNUC2A	large intestine
COV504	ovary	PANC0813	pancreas
HCC70	breast	RKO	large intestine
NCIH508	large intestine	SU8686	pancreas
TE15	esophagus	GP2D	large intestine
A204	soft tissue	SW48	large intestine
QGP1	pancreas	COV434	ovary
TT	esophagus	HUTU80	small intestine
JHOC5	ovary	LN229	cns
SNU840	ovary	RKN	ovary
CAOV4	ovary		

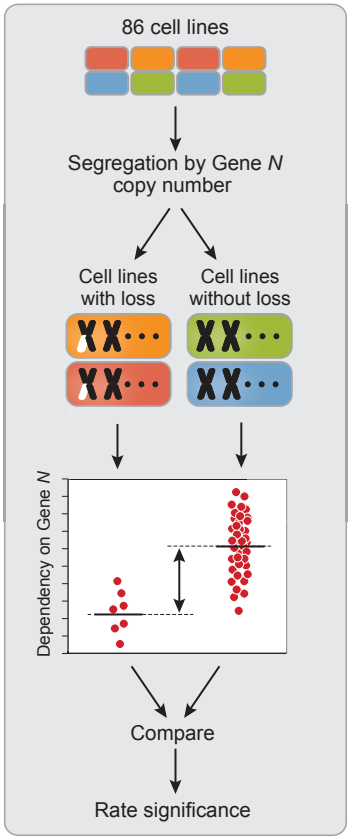


Figure 11- Analysis to find cancer vulnerabilities predicted by partial loss: Schematic describing the approach to identifying CYCLOPS genes. For each gene, we separated cell lines with and without loss of the gene and compared their dependency on that gene by permuting class labels.

Table 7 - Top Genes from CYCLOPS analysis

CYCLOPS Analysis of 5312 genes (6085 Gene Dependence Solutions)				
Gene	cyto band	# of loss lines	# of neutral lines	p value
PSMC2	7q22.1	11	73	0.00002
EIF2B2	14q24.3	17	68	0.00002
EEF2	19p13.3	35	48	0.00002
PHF5A	22q13.2	24	62	0.00002
HPGD	4q34.1	34	48	0.00004
RPS15	19p13.3	33	50	0.00004
SNRPB	20p13	10	74	0.00004
POLR2F	22q13.1	22	63	0.00004
USPL1	13q12.3	27	58	0.00008
SMC2	9q31.1	19	67	0.00012
SMU1	9p13.3	27	57	0.00014
PUF60	8q24.3	10	75	0.00016
RPS11	19q13.33	16	70	0.0002
POLG	15q26.1	18	67	0.00022
ZNF583	19q13.43	12	74	0.00022
CPT1B	22q13.31	23	61	0.00022
BMP8A	1p34.2	9	77	0.00024
TIE1	1p34.2	7	78	0.00028
SF3A2	19p13.3	34	49	0.00028
SNRNP70	19q13.33	17	69	0.0003
RBM17	10p15.1	21	60	0.00032
PCNA	20p12.3	10	73	0.00032
PSMA4	15q25.1	20	64	0.00036
LSM4	19p13.11	21	62	0.00036
EEF1A1	6q13	20	66	0.00048
FBXO6	1p36.22	13	73	0.00052
ASCL3	11p15.4	30	53	0.00056
PGF	14q24.3	17	68	0.00056
ETV2	19q13.12	12	74	0.00056
PAFAH1B1	17p13.3	31	55	0.00068
UBA52	19p13.11	20	63	0.00068
OBP2A	9q34.3	11	74	0.00074
PABPN1	14q11.2	18	67	0.00076
NUPL1	13q12.13	29	55	0.00078
CEBPG	19q13.11	14	72	0.00082
PSMC4	19q13.2	10	75	0.0009
PARK7	1p36.23	15	71	0.00096

To minimize the confounding effect of lineage, all permutations maintained the initial lineage distribution within each class. We also restricted these analyses to the 5,312 genes for which each class contained at least seven cell lines. We identified 56 candidate genes with False Discovery Rate (FDR)[192] p-values less than 0.25 (**Table 7**) and named them “CYCLOPS” genes (Copy-number alterations Yielding Cancer Liabilities Owing to Partial losS).

Recently, RNAi pooled screens have come under fire for the discordance between individual experiments, as well as potential domination of signal from off-target effects relating to miRNA[313]. To assess how much of our CYCLOPS signal could be attributed to on-target activity, we looked to see how the correlation between a gene’s dependency score and copy number compared to correlations with copy number of other genes. We found significantly more genes with positive correlation between copy number and dependency score than when we performed the same test with scrambled gene labels (Pearson’s correlation, 71 observed genes with FDR $p<0.25$, scrambled set mean = 1.42, sd = 1.91, $p<0.0001$) This suggests the correlation between copy number and gene dependency we observe is unlikely to be observed by chance, if, for example, our data was driven solely by off-target effects. In addition, we validated the CYCLOPS vulnerabilities using an alternative RNAi dataset (shRNA Activity Rank Profile, or shARP)[312] representing consequences of expressing 78,432 shRNAs targeting 16,056 genes on the proliferation of 72 breast, ovarian, or pancreatic cancer cell lines. We applied the same analysis pipeline, constrained to the “Validation Set” of 47 cell lines for which we had copy-number data and the 6,574 genes for which at least seven cell lines were in each class (copy-loss and copy-neutral). These genes included 3,282 of the

genes that underwent full analysis in the Achilles dataset and 40 of the CYCLOPS candidates identified in that analysis. Although the lineage distribution was markedly different between the Validation and Achilles datasets (breast and pancreatic cancers made up 90% of the cell lines in the Validation set but only 15% in Project Achilles), the 40 CYCLOPS candidates identified in the Achilles analysis tended also to be highly enriched for relative sensitivity after partial loss in the shARP analysis (KS statistic, $p=2x10^{-9}$).

Features of CYCLOPS genes

In copy-number analyses collected from 3,131 tumor samples and cancer cell lines[314], each CYCLOPS candidate was found to undergo hemizygous loss in an average of 18.5% of samples (range 8%-33%), which was as common as for the other 5,256 genes in the analysis (average 17.7%, range 4%-34%; two tail p=0.17). In contrast, CYCLOPS genes exhibited much lower rates of homozygous deletion (p=0.02) and DNA methylation (p=0.0045) (**Fig. 12a**). This observation suggested that CYCLOPS genes are enriched for genes required for cell proliferation or survival. Indeed, we used the set of genes found to be essential in S. cerevisiae[315] to identify 1,336 homologous human genes and found that CYCLOPS genes are not only highly enriched (p<0.0001) in this group of essential genes but show comparable rates of genetic and epigenetic alterations (**Table 8**).

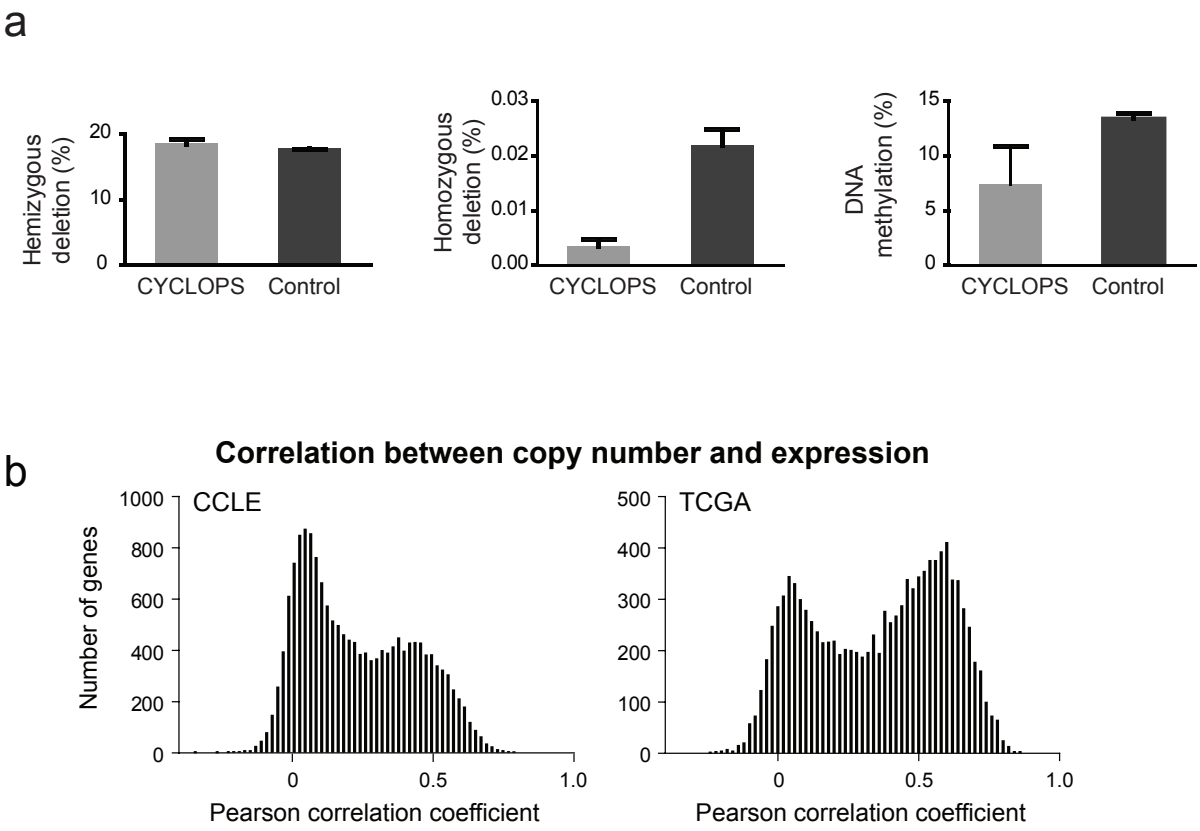


Figure 12- Characteristics of CYCLOPS genes: (**a**) Frequency of hemizygous deletion, homozygous deletion, or DNA methylation of CYCLOPS and other genes. Data are presented as averages ± S.E.M. (**b**) Strength of correlation between copy number and mRNA expression level across genes profiled in the CCLE and TCGA ovarian cancer datasets. Results represent Pearson correlation coefficients between $\log_2$ copy number levels and quantile normalized mRNA levels for 16,767 genes across 806 cell lines in the CCLE (Barretina et al., 2012) and 11,119 genes across 429 ovarian tumors from the TCGA (Bell et al., 2011). (**c**) Strength correlation between copy number and expression in genes categorized by their expression across multiple tissues. Genes were separated into two categories based on expression patterns in 6 immortalized cell lines, as determined by RNA-seq. Genes with RPKM vales were greater than 0.1 for all samples were considered “ubiquitously expressed” (blue) versus all other genes assayed were (red).

Table 8 - rates of loss in essential genes		
Copy Number and Methylation Frequencies for Essential Genes		
Essential Genes Analyzed (n=1277)		
Control set (n=20,597)		
	Essential	Control
Homozygous Deletion Frequency	0.0077% +/- .00168%	0.015798% +/- 0.00122%
Partial Genomic Loss Frequency	15.63% +/- 0.19%	15.74% +/- 0.04765%
Essential Genes Analyzed (n=963)		
Control set (n=18800)		
Hypermethylation	5.59% +/- 0.64%	11.96% +/- 0.2%

A pathway enrichment analysis of CYCLOPS candidates showed the spliceosome, proteasome, and ribosome were the most highly enriched pathways (KS statistic FDR p = $1.4x10^{-8}$, $2.7x10^{-5}$, and $1.8x10^{-4}$ respectively, **Table 9**, **Methods**). These pathways were also the most significantly enriched in our analysis of the Validation set (spliceosome FDR=$3.1x10^{-15}$; proteasome FDR=$1.5x10^{-12}$; ribosome FDR=$2.3x10^{-17}$). Taken together, these observations indicate that CYCLOPS genes are a unique subset of cell essential genes for which partial but not complete suppression is compatible with cancer cell viability.

Table 9 - KEGG pathway enrichment among CYCLOPS candidates

KEGG pathway	FDR value	# of CYCLOPS genes
Spliceosome	1.40E-08	9
Proteasome	2.70E-05	3
Ribosome	1.80E-04	4

These observations lead us to hypothesize that copy-number loss might unveil vulnerabilities in CYCLOPS genes through decreased gene expression. We therefore evaluated the relation between copy-number loss and expression using integrated SNP

and expression data for 16,767 and 11,118 genes respectively in two panels of samples: the Cancer Cell Line Encyclopedia (CCLE) of 805 cell lines across 24 cancer types[316], and 429 ovarian cancers profiled by The Cancer Genome Atlas Project[317]. In consonance with prior reports from other datasets[318], the correlation between copy-number and expression revealed that genes fall into one of two classes: a class in which mRNA levels are relatively independent of copy-number (CCLE: modal r = 0.05; TCGA: modal r = 0.07), and a second class in which copy-number and gene expression are more closely correlated (CCLE: modal r = 0.49; TCGA: modal r = 0.61; **Fig. 12b**). We found that CYCLOPS candidates were enriched in the latter class (CCLE p = 0.0004; TCGA p = 0.04), and indeed the average strength of correlation between copy loss and mRNA expression was significantly higher for CYCLOPS candidates than for the other genes in our analysis (CCLE r = 0.39 vs 0.26, $p < 0.0001$; TCGA, r = 0.44 vs 0.34, p = 0.0017). Notably, amongst all 158 KEGG pathways, the three pathways most enriched in CYCLOPS candidates also have significant correlations between copy-number and expression: spliceosome (CCLE r = 0.46, FDR=2.2×10^{-5}; TCGA r = 0.56, FDR = 3.9×10^{-5}), proteasome (CCLE r= 0.52, FDR= 2.2×10^{-5}; TCGA r= 0.60, FDR= 3.9×10^{-5}) and ribosome (CCLE r=0.44, FDR=2.2×10^{-5}; TCGA r=0.47, FDR=1.1×10^{-4}).

This bimodal distribution of associations between copy-number and expression appears to reflect two classes of genes: those that are ubiquitously expressed, and those that are selectively expressed. We used RNA-seq data from 7 cell lines across multiple lineages (GM12878 Lymphoblasts, K562 AML,H1HESC human embryonic stem cells, HSMM_human skeletal muscle, HUVEC human vascular epithelium, HMEC breast normal, HCC1954 breast cancer) and classified genes as either being

ubiquitously expressed (RPKM >0.1 for all cell lines) our selectively expressed (RPKM for at least one cell line). We favored RNA sequencing over array based expression platforms because the former are more reliable for comparisons across genes[319]. The ubiquitous vs. selective expression classification is strongly associated with the bimodal distribution we observe in correlation between copy number and expression in the CCLE (t-test p=0; **Fig. 12c**). These observations indicate that mRNA expression of essential and ubiquitously expressed genes, and CYCLOPS genes in particular, is reduced as a result of copy loss.

Discussion

Discovery of new vulnerabilities due to genomic disruption in cancer

Advances in cancer therapeutics benefit from our ability to identify vulnerabilities predicted by genomic features that are unique to cancer cells. Indeed, the inhibition of recurrent activating mutations in proto-oncogenes has led to several new cancer treatments. The cancer-specific vulnerabilities we have identified herein are the consequence of alterations in genes affected by genomic disruption that may have no consequences to the process by which the cell transformed or continues to proliferate. These genomic alterations are more frequent than most known driver alterations, occur across lineages, and could theoretically be targeted in a large number of patients.

CYCLOPS genes as synthetic lethal targets

CYCLOPS genes represent a specific form of synthetic lethality. Several studies have investigated synthetic lethality with activation of pathways that drive cancer but that cannot themselves be easily targeted. For example, synthetic lethality is one approach to targeting inactivated tumor suppressor genes, whose functions cannot

easily be reconstituted. This approach has had success in breast and ovarian cancers that have *BRCA1* or *BRCA2* loss, and as a result lack the homologous recombination DNA repair pathway, making them exclusively depend on the nucleotide excision DNA repair pathway (Bryant, 2005; Farmer, 2005). Targeting CYCLOPS genes represents a different approach to synthetic lethality, in which the intervention is synthetic lethal with a genetic event independent of the effects that event has on the pathways that drive cancer.

Cancer heterogeneity, evolution, and passenger alterations.

One concern raised with targeting passenger alterations is the emergence of resistant clones and their implications on long-term therapeutic efficacy. Unlike driver alterations, cancers are not dependent on the specific alteration for tumor maintenance, but instead are vulnerable based on their specific evolutionary trajectory. One may imagine that overcoming this vulnerability would be easier than finding new ways to replicate the function of targeted driver alterations. While this assumption is reasonable, it is still worthwhile to pursue expansion of therapies to non-driver targets.

Chapter 5: The proteasome and PSMC2 as a Cyclops target

Section goals

Having isolated a set of potential cancer vulnerabilities predicted by partial loss in a panel of cancer cell lines, the goal of this section was to determine the mechanism of vulnerability for one of these candidates. We chose *PSMC2* for two reasons. First, it was the highest-ranked CYCLOPS candidate in our original analysis, and was also significant in the validation RNAi screen[320]. Second, it is a member of the proteasome, whose components were more generally enriched among CYCLOPS genes, and which already serves as a therapeutic target in cancer.

The proteasome is the primary complex responsible for protein degradation. Proteins are constantly being produced and degraded in living cells[321], and much of the early work on protein degradation elucidated energy independent enzymes such as trypsin and other proteases. However, experiments using labeled amino acids in the 1950s discovered that the majority of amino acids released from organ-derived samples was inhibited under anaerobic conditions and cyanide addition, suggesting that this was an energy dependent process[322]. While it was understood that protein turnover is a constant process, Goldberg discovered that puromycin-induced early termination of polypeptides led to their being more rapidly degradation[323], suggesting that protein folding can affect in vivo stability[324]. The proteasome was the primary organelle involved in this process, and was found to be an ATP-dependent protease[325] responsible for both mis-folded protein degradation[326] and regulation of active protein levels through programmed degradation. This is directed by covalently linking targeted

proteins with a small, repeat peptide named ubiquitin, a process whose elucidation was awarded the Nobel prize[327,328].

The term “proteasome” applies to a set of similar, large (over 2.5 megadaltons) complexes all involved in protein degradation. The 28 proteins in common to all proteasome assemblies are collectively referred to as the 20S proteasome, to which multiple independent regulatory complexes may be attached, each imparting different functionality[329]. Production of this large and intricate machine takes place in a highly coordinated and reproducible procedure, with the construction and attachment of each subcomplex occurring in a precise order[330-333]. Additionally, because individual proteins non-functional in isolation, and are required in specific stoichiometric amounts during complex formation, most of the subunits are under tight transcriptional control through a negative feedback loop involving a single transcription factor[334-337]. While this coordination is presumably efficient in regulating levels of proteasome in response to cellular requirement, it may make it more challenging for the cell to respond to deficiency in a single subunit of the complex, such as that generated by genetic loss in cancer.

The 19S proteasome is the most abundant subassembly to co-complex with the 20S proteasome, forming the 26S proteasome, which is the fundamental complex responsible for ubiquitin mediated protein degradation[338]. *PSMC2* is a ATPase subunit of the 19S complex that catalyzes the unfolding and translocation of substrates into the 20S proteasome[330].

Direct inhibition of the proteasome is a vulnerability in multiple myeloma due to the specific cellular requirements of these cells. Multiple myeloma is a malignancy

derived from plasma cells and their aberrant production of a large amount of immunoglobulin may lead to heavy reliance on the unfolded protein response (UPR) system[339-341]. Even though proteasome function is necessary for eukaryotic cellular survival, proteasome inhibitors, which block the UPR pathway and lead to apoptosis, were found to be well-tolerated in clinical trials[342] and bortezomib is now considered a first-line therapy for multiple myeloma[343]. The clinical success of bortezomib, which targets an essential cellular component, opens the doors for other therapies that look for a therapeutic window between cancer and normal cells, even in pan-essential pathways.

Methods

Cell Culture

Cancer cell lines were obtained from the Cancer Cell Line Encyclopedia. The generation of immortalized ovarian surface epithelial cells was previously described[344]. TYKNU and HEYA8 were cultured in DMEM supplemented with 10% Fetal bovine serum and 2mM L-Glutamine and all the remaining cell lines were cultured in RPMI 1640 supplemented with 10% Fetal bovine serum.

Lentiviral production

Lentiviruses were produced for expression constructs or shRNA constructs in 293T cells cultured in DMEM with 10% FBS supplemented with 4 mM L-Glutamine using the three-vector system as described[344]. The virus was diluted (1:10) and added to 2 x 10^5 cells in a 6 well plate containing 8 µg/ml of polybrene (Sigma). Plates were centrifuged for 15 min, 1126 X *g* at room temperature. For selection of virally infected cells 24 h post infection 2 µg/ml of Puromycin (LKO.1) or 10 µg/ml of Blasticidin

(LEX303) was added. Where applicable, lysate was collected and protein levels were analyzed 5 d after infection.

Competition Assays using shRNA constructs

We found that a single lentiviral integrant expressing either shRNA-1 or shRNA-3 was sufficient to suppress PSMC2 levels in *PSMC2*^{Neutral} cells relative to cells that express shLacZ (**Fig. 13c**); competition assays were therefore performed at a multiplicity of infection of 1. To perform competition assays, *PSMC2* shRNA-3, *PSMC2* shRNA-4, and *PSMC2* shLacZ in pLKO.1 were modified by inserting *GFP* into the puromycin cassette (BamHI/Kpn1) to yield pLKO GFP constructs that expressed these shRNAs. Cells were infected with lentivirus that contained the indicated shRNA in pLKO.1 GFP and treated for 24 h with three fold dilutions of virus according to protocols for lentiviral infection. At 48 h, the cells were analyzed using a BD LSR II flow cytometry system (BD Biosciences) for GFP+ cells. The well in which the viral titer resulted in approximately 50% of GFP+ cells was then cultured for 21 d. Using FACS, we analyzed these cell populations for GFP+ cells on days 7, 14, and 21 and recorded the percentage of GFP+ cells normalized to the day 0 time point.

PSMC2 shRNA inducible system

The following sense-antisense oligonucleotides (IDT) were annealed and then cloned into Tet-pLKO-neo (Addgene #21916) (AgeI/EcoRI).

shRNA	Sense	Antisense
PSMC2 shRNA-2	5'CCGGGCCAGGGAGATTGGATAGAAACTC GAGTTTCTATCCAATCTCCCTGGCTTTTTG	5'AATTCAAAAAGCCAGGGAGATTGGATAGAAA CTCGAGTTTCTATCCAATCTCCCTGGC
PSMC2 shRNA-3	5'CCGGGCCTGCCTTATCTTCTTTGATCTCG AGATCAAAGAAGATAAGGCAGGCTTTTTG	5'AATTCAAAAAGCCTGCCTTATCTTCTTTGATCT CGAGATCAAAGAAGATAAGGCAGGC
PSMC2 shLacZ	5'CCGGTGTTCGCATTATCCGAACCATCTC GAGATGGTTCGGATAATGCGAACATTTTTG	5'AATTCAAAAATGTTCGCATTATCCGAACCATC TCGAGATGGTTCGGATAATGCGAACA

Plasmid sequences were confirmed by sequencing. Lentivirus generated using these constructs was used to infect either OVCAR8 or A2780 cells. Stably infected cells were selected with Geneticin (500 ug/ml) (Sigma). Cells were treated with 100 ng/ml of doxycycline unless otherwise indicated (Sigma). Beginning 4 d after the addition of doxycycline, OVCAR8 cells begin to die. Therefore, to assess the level of PSMC2 suppression in OVCAR8 cells, we collected cells 3 d after the addition of doxycycline. Since the proliferation and viability of A2780 cells, in contrast, are not affected by the addition of doxycycline, we collected cells 4 d after the addition of doxycycline at which point PSMC2 levels had achieved a new steady state. All proliferation or viability studies on both cell lines were conducted 7 d after the addition of doxycycline. Lysates from cells using this system were made in Buffer A for all immunoblots except for Ubiquitin immunoblots, which were made in Buffer B.

Ectopic V5-PSMC2 expression

PSMC2 was engineered with a V5 sequence at the N terminus and cloned into pLEX303. To minimize a second translation initiation site that used the endogenous ATG, we amplified PSMC2 using a 5' primer that contained a sequence that would code for the V5 epitope and also mutated the endogenous methionine to a threonine. The sequence of the exogenous construct added the following amino acids to the N-

terminus (MGKPIPNPLLGLDST) where the final T (Threonine) is in place of the endogenous methionine. No modifications were made to the C-terminus. OVCAR8 or IOSE cells were serially infected with lentivirus containing either pLEX303 GFP (obtained from the TRC) or pLEX303 V5-PSMC2. Repeat infections were performed with respective constructs until the expression of ectopic V5-PSMC2 (as measured by western blot) was comparable to the endogenous protein. OVCAR8 cells that expressed either V5-PSMC2 or GFP were infected with lentivirus that express PSMC2 shRNA-1 or PSMC2 shLacZ in pLKO.1. 5 d after infection, the cells were analyzed by immunoblot and 7 d after infection, in triplicate the cells were analyzed for total ATP content (Promega). Relative ATP content was normalized to the cells infected with shLacZ. Finally, cell lines used for orthotopic xenograft tumors were engineered to express firefly luciferase. Lentivirus made from vectors containing the luciferase gene (LEX301) and obtained from the TRC was used to infect the indicated cell lines.

Cellular Protein Lysate

All cells were first harvested and pelleted in cold PBS. All subsequent procedures were performed at 4°C. For lysate generated in “Buffer A”, the cell pellet was resuspended in 10% Glycerol, 25mM Hepes pH 7.4, 10mM $MgCl_2$, 1mM ATP, 1mM DTT, and phosphatase and protease inhibitors without EDTA (Roche). Sonication was performed at low intensity using a micro-tip on ice for 1 min (50% cycle). The resulting cell mixture was centrifuged at 13000 X *g* for 15 min at 4°C. The supernatant was collected and centrifuged at 100,000 X *g* for 60 min. The subsequent supernatant was collected and used in future studies as lysate. Protein amount was normalized using the Bradford reagent (Bio-Rad). For lysate generated in “Buffer B”, the cell pellet was

resuspended in 20 mM Tris-HCL pH 7.5, 150mM NaCl, 10% Glycerol, 1% Triton-X, 0.2 mM DTT, 250 ug/ml NEM supplemented with protease inhibitors (Roche) and Phosphatase Inhibitors (PhosStop -Roche). Samples were incubated for 15 min and supernatants were collected after centrifugation at 15,000 X *g*. Finally, for lysate generated in "RIPA", the cell pellet was resuspended in 1X RIPA buffer (Boston Bioproducts) that was supplemented with protease inhibitors with EDTA (Roche) for 15 min. The supernatant was collected after centrifugation at 15,000 X *g* for 15 min. Protein lysates made in RIPA or Buffer B were normalized with THERMO BCA normalization kit (Bio-Rad), using a BSA standard curve. Lysate analysis of 26S proteasome components across ovarian cancer cell lines and PSMC2 levels following shRNA expression were generated in RIPA buffer. For all experiments that involved native analysis of the proteasome, lysate was made in Buffer A and then analyzed either in non-denaturing or denaturating (SDS loading buffer) conditions.

Purification of Complex^{PSMC2} and the 26S proteasome

Continuous glycerol gradients (from 10%-40%) were made in 25mM Hepes pH 7.4, 10mM $MgCl_2$, 1mM ATP, 1mM DTT. 18 mg of lysate (Buffer A) was loaded at the top of a 14 ml gradient and centrifuged for 20 hours at 195,000 x *g*. Fractions were removed from top of each gradient in 1 ml increments. Native PAGE described above was used to determine which fractions contained Band A or 26S proteasome and the indicated pooled fractions were then pooled and incubated with Anti-V5 agarose conjugates. Immunoprecipitates were eluted and analyzed by immunoblot.

26S proteasome activity

In vitro, we measured excitation-emission spectra (360 nm to 430 nm) during incubation at 37°C every 30 sec for 1 h for a 100 µl solution containing 5 µl of lysate (Buffer A) in 50mM Tris-HCl pH 8.0, 40mM KCl, 5mM $MgCl_2$, 1mM ATP, 1mM DTT, and 100µM Sucrose-LLVY-AMC (Bachem). We converted these measurements to amount of peptide cleavage using a standard curve generated from the excitation-emission spectra of AMC (Bachem). Samples were tested in triplicate with and without the addition of 1 µM bortezomib. The average value of peptide cleavage in the bortezomib sample was subtracted to determine 26S proteasome activity.

Native gel analysis for proteasome content or proteasome activity

10 µg of lysate (Buffer A) was loaded onto 3-8% Tris-Acetate PAGE (Invitrogen) and run in Tris-Glycine at 4°C and 60V for 17 h. Gels were transferred to nitrocellulose membranes in Tris-Glycine at 70V for 4 h for immunoblotting or in gel peptidase activity. The latter was performed by incubating with gentle agitation in 50mM Tris-HCl pH 8.0, 5mM $MgCl_2$, 1mM ATP, 1mM DTT, and 50µM Suc-LLVY-AMC (Bachem) at 37°C for 30 min. Gels were visualized under UV transillumination. Following photography of 26S proteasome activity, gels were incubated for another 45 minutes at 37°C in the same buffer with the addition of 0.2% SDS and re-analyzed by UV transillumination to assess 20S peptidase activity.

Immunoprecipitation of Complex$^{PSMC1/PSMC2}$ and the 26S proteasome

Immunoprecipitation was performed by incubating fractions with 50 µl of Anti-V5 agarose conjugated beads (Sigma) and rotating for 16 h at 4°C. Each wash step began with centrifugation at 1000 X *g* for 3 min, removal of the supernatant, and resuspension in 1 ml of buffer A. Three washes were completed and the samples were then incubated

in 100 µl of 20 µM V5 peptide (Sigma) in buffer A. Equal volumes were then denatured in SDS loading buffer and analyzed by immunoblot.

PSMC2 mRNA levels determination

Cells were harvested and pelleted in PBS. RNA was extracted from cell pellets and resuspended in Trizol Reagent (Invitrogen) according to manufacturer's protocol. The precipitated total RNA was then resuspended in 0.1 ml of a 1x TurboDNase buffer with 2U of DNase (Ambion) and incubated at 37°C for 15 min. RNA was then purified using the Qiagen RNEasy kit according to manufacturer's instructions for "RNA Clean Up". 1 µg of RNA was used to generate cDNA using the M-MLV Reverse Transcriptase kit (Invitrogen) according to the recommended protocol. PCR reactions were performed in replicates of five using SybR PCR master mix (Applied Biosystems) and Ct values were automatically determined using Applied Biosystems 7300 System software. The resulting data were normalized to housekeeping genes and analyzed using the delta-delta Ct method for fold difference between control and test samples. Unless otherwise indicated *PGK1* was used as an internal control. Primers used in quantitative RT-PCR are as follows: *PGK1* (5'AGAGGGAGCCAAGATTGTCA, 5'GGTATGCCAGAAGCCACAGT), *Tubulin* (5'TCTGTTCGCTCAGGTCCTTT, 5'TGTGTCCTTGCACCCAAATA), and *PSMC2* (5'TCCACCCGGTACAGGCAAGACACT, 5'CGAGCCCCCTCACCGACGTA)

***PSMC2* siRNA experiments in A2780 cells**

5 x 10^5 cells were plated in 10 cm plates on Day 0. Each plate was transfected with a total of 500 pmol of siRNA and 20µL lipofectomine RNAi Max (Invitrogen) using the manufacturer's recommended procedure. Three pre-annealed PSMC2 siRNAs were

obtained from IDT DNA and pooled in equimolar ratios: siRNA-1 (5'-GCUGUAAAUAAGGUCAUUAAGUCTT, 5'-AAGACUUAAUGACCUUAUUUACAGCUU), siRNA-2 (5'-AGAUAAUCAAUGCUGAUUCGGAGGA, 5'-UCCUCCGAAUCAGCAUUGAUUAUCUUU), siRNA-3 (5'-CCCACAUAUUUAAGAUUCACGCUCG, 5'-CGAGCGUGAAUCUUAAAUAUGUGGGUC). The concentration of the pool was varied and the difference was made up with control siRNA (Ambion). 24 hours after transfection, 2.5 x 10^5 viable cells (determined by trypan blue) were plated in duplicate for a six-day proliferation assay. The remaining cells were plated and harvested for lysate (RIPA) 3 d after transfection.

***PSMC2* siRNA experiments in A2780-Dox cells**

1.5 x 10^6 of A2780-Dox-shLacZ and A2780-Dox-shRNA-2 cells were plated in 10 cm plates on Day 0, two for each cell line. On Day 1, the media from one plate of each cell line was replaced with media containing 100 ng/ml of doxycycline. On day 2, these cells were passaged into 96 well plates at 2,500 cells/well, as well as a single 10 cm plate (to be used for RNA) for each condition and cell line. On Day 3, these cells were transfected with 0.2 mM lipofectomine RNAi Max (Invitrogen) along with siRNA at a total concentration of 20 nM, with varying levels of siRNA specific for either *PSMC2* or *PSMC5* using the manufacturer's recommended procedure. *PSMC2* siRNA are listed above, with additional *PSMC5* siRNAs obtained from IDT DNA predesigned siRNA (Catalog # HSC.RNAI.N002803.12.1, HSC.RNAI.N002803.12.2, HSC.RNAI.N002803.12.3) and pooled in equimolar ratios. The concentration of the pool was varied and the difference was made up with control siRNA (Ambion). Proliferation was measured by cell titer glo, and qRT-PCR data was used to determine concentrations where siPSMC2 (2.5 nM) and siPSMC5 (5 nM) led to similar levels of suppression of their respective targets.

PSMC2 and Control siRNA nanoparticle siRNA sequences

Gene	Sense Strand	Antisense Strand
siRNA-1	5'- GCUGUAAAUAAGGUCAUUAUU	5'- UAAUGACCUUAUUUACAGCUU
siRNA-2	5'- GCCAGGUGUACAAAGAUAAUU	5'- UUAUCUUUGUACACCUGGCUU
siRNA-3	5'- GGACCCACAUAUUUAAGAUUU	5'- AUCUUAAAUAUGUGGGUCCUU
GFP	5'-GGCUACGUCCAGGAGCGCA	5'-UGCGCUCCUGGACGUAGCC

Bortezomib sensitivity experiments

Throughout the experiment, A2780 cells engineered in the PSMC2 shRNA inducible system were either treated with vehicle or 30 ng/ml of doxycycline. Doxycycline treatment was started on Day 0. On day 4, cells were plated at 2000 cells/ well in a 96 well plate. The following day, the cells were treated with varying concentrations of bortezomib or vehicle. Total ATP levels were measured by Cell Titer Glo (Promega) 72 h after adding bortezomib. The data was normalized to the vehicle treated sample. Graphpad was used to determine the IC_{50} by constructing a non-linear regression with a four-parameter variable slope.

Cell Cycle and Apoptosis Assays using the PSMC2 shRNA inducible system

We cultured either A2780 or OVCAR8 cells engineered with our PSMC2 shRNA inducible system in the presence or absence of doxycycline (100 ng/ml) and collected them for analysis after 3 d. We used the BrdU Flow Kit (BD Pharmigen) according to manufacturer's protocol to determine the percentage of cells in each phase of the cell cycle 72 h after the addition of doxycycline. At the same time point, we independently determined the number of cells undergoing apoptosis by FACS analysis of Annexin-5 according to the manufacturers recommended procedure (Invitrogen).

Generation of PSMC2 and Control siRNA nanoparticles

The generation of tumor-penetrating nanocomplexes carrying PSMC2-siRNA (Dharmacon) and measurement of their uptake and effects on cellular proliferation was

performed as described[345]. The p32-receptor specificity of cellular uptake was probed by applying a monoclonal antibody directed against p32 (100 ug/mL) to cells 1 h prior to the addition of TPN. More information about the reagents, chemicals and siRNA sequences can be found in the Extended Experimental Procedures.

Generation of orthotopic xenografts and nanoparticle administration

10^6 OVCAR8 cells, 0.5×10^6 OVCAR8 cells expressing V5-PSMC2, or 0.2×10^6 A2780 cells expressing doxycycline-inducible shRNA against *PSMC2* were implanted intraperitoneally in 4-6-week-old NCr/nude mice (Charles River). Once tumors were established and confirmed by bioluminescence imaging, animals were treated intraperitoneally with nanoparticles carrying *GFP*-specific siRNA (TPN/siGFP), or TPN containing *PSMC2*-specific siRNA (1 mg siRNA/kg/injection) every 3 d for 21 d as described[345]. Mice bearing A2780 tumors expressing the doxycycline-inducible sh*PSMC2* were continuously fed with doxycycline-containing diet (2000 mg/kg) beginning two days after tumor cell injection. Mice were sacrificed and tumors harvested at the end of the experiment or when the tumor burden resulted in a failure to thrive according to institutional recommendations. Tumor lysates were made by homogenizing tumors using an eppendorf micropestle in RIPA buffer supplemented with protease inhibitors. We restricted analyses of PSMC2 expression to tumors that were relatively devoid of mouse tissue.

Author Contributions

The author, in collaboration with Deepak Nijhawan and Matt Strickland, were responsible for viral production, transfections, infections, in vitro western blots and cell culture maintenance.

The author was responsible for native gel western blotting and peptidase cleavage, siRNA experiments, qPCR analysis, glycerol gradients, and Immunoprecipitation of proteasome complexes, with assistance from Deepak Nijhawan and Matt Strickland.

The author and Rebecca Lamothe were responsible for western blots and qPCR on samples derived from in vivo models.

Results

To test the possibility that our observations were the result of a confounding genetic alteration, we determined whether expression and copy-number levels of every other gene for which we had expression or copy-number data significantly correlated with *PSMC2* sensitivity. Low *PSMC2* expression (FDR < 0.017) and *PSMC2* copy loss (FDR < 0.008) were the features most significantly correlated with PSMC2 sensitivity genome-wide. Conversely, among the 7,250 genes in our Achilles analysis, only sensitivity to *PSMC2* correlated with *PSMC2* copy loss (FDR < 0.25). In particular, amongst all 47 other proteasome components surveyed, neither expression levels nor copy-number status significantly correlated with *PSMC2* sensitivity. Suppression of the 29 proteasome components in the Achilles data also did not specifically inhibit proliferation of cell lines with *PSMC2* copy loss. The association between *PSMC2* copy loss and PSMC2 sensitivity also remained significant when cells with PSMC2 copy-number gains were excluded from the analysis (p=0.0006).

Since partial copy loss of cell essential genes, like *PSMC2*, might afford only small differences in sensitivity to suppression between different cells, we also compared

the effects of *PSMC2* suppression to that observed when we suppressed the oncogenes *KRAS*, *PIK3CA*, and *BRAF*. These oncogenes are associated with some of the most specific known cancer dependencies through "oncogene addiction"[346]. In consonance with prior studies, suppression of these oncogenes inhibited proliferation of cells harboring mutated and constitutively active oncogenes compared to cells expressing wild type proto-oncogenes ($p<2x10^{-5}$ in each case) (**Fig. 13a**). However, the difference in *PSMC2* dependency scores between cell lines with and without *PSMC2* copy loss (*PSMC2*^{Loss} and *PSMC2*^{Neutral}, respectively) was greater than for any of these three models of oncogene addiction (**Fig. 13a**).

We confirmed the vulnerability of *PSMC2*^{Loss} lines to PSMC2 suppression in a direct competition assay by comparing the proliferation rate of uninfected cells to cells that co-express GFP and either shLacZ or a *PSMC2* shRNA in six ovarian cell lines over 21 days. The expression of shLacZ or *PSMC2* shRNAs failed to induce significant changes in the proliferation of *PSMC2*^{Neutral} cells, including two ovarian cancers and one non-transformed Immortalized Ovarian Surface Epithelial cell (IOSE) cell line[344]. In contrast, expression of *PSMC2* shRNAs reduced the proliferation rate by at least 50% in all three *PSMC2*^{Loss} ovarian cancer cell lines within 7 days (**Fig. 13b-d**).

To confirm that these observed effects were due to the suppression of PSMC2, we expressed an N-terminal V5-epitope tagged form of *PSMC2* (hereafter referred to as *V5-PSMC2*) in OVCAR8, a *PSMC2*^{Loss} cell line. *V5-PSMC2* expression was unaffected by an shRNA that targets the 3' UTR of endogenous *PSMC2*, and rescued the proliferation of OVCAR8 cells that express this shRNA (**Fig. 13d**).

PSMC2 levels and survival in *PSMC2*^{Loss} cell lines

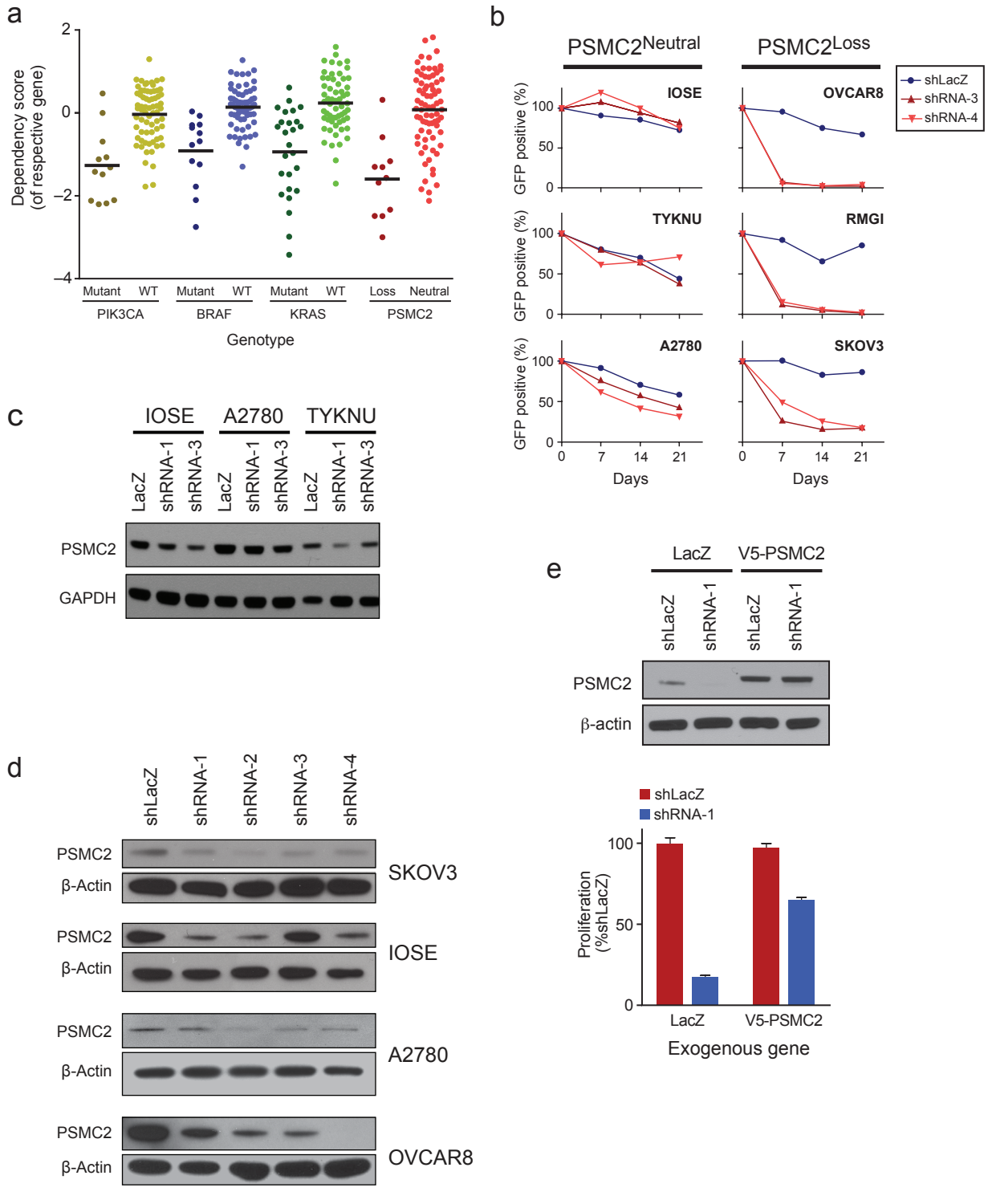

a
Dependency score (of respective gene)
2
0
−2
−4
Mutant
WT
Mutant
WT
Mutant
WT
Loss
Neutral
PIK3CA
BRAF
KRAS
PSMC2
Genotype
b
PSMC2Neutral
PSMC2Loss
shLacZ
shRNA-3
shRNA-4
IOSE
OVCAR8
TYKNU
RMGI
A2780
SKOV3
GFP positive (%)
100
50
0
0
7
14
21
Days
c
IOSE
A2780
TYKNU
LacZ
shRNA-1
shRNA-3
PSMC2
GAPDH
d
shLacZ
shRNA-1
shRNA-2
shRNA-3
shRNA-4
PSMC2
β-Actin
SKOV3
IOSE
A2780
OVCAR8
e
LacZ
V5-PSMC2
shLacZ
shRNA-1
PSMC2
β-actin
shLacZ
shRNA-1
Proliferation (%shLacZ)
100
50
0
LacZ
V5-PSMC2
Exogenous gene

Figure 13- PSMC2 as a CYCLOPS gene: (**a**) Comparison of gene dependence between three models of oncogene addiction and *PSMC2*. Cell lines were classified by mutation status for *PIK3CA*, *BRAF*, or *KRAS* (n=102 in each case) or *PSMC2* copynumber (n=84). For each class, gene dependency scores reflect the sensitivity to the gene on which the categorization was based. Solid bars represent average scores. (**b**) The effect of *PSMC2* suppression on the proliferation of six ovarian cell lines. (**c-d**) *PSMC2* suppression in ovarian cell lines. We individually expressed all four *PSMC2* shRNAs and a control shRNA, shLacZ. Among the four *PSMC2* shRNAs, *PSMC2* shRNA-2, 3, and 4 were used by ATARiS to calculate the *PSMC2* dependency score and exhibited consistent suppression of PSMC2. (**e**) PSMC2 levels (top) and relative proliferation rates (bottom) among cells expressing different combinations of *PSMC2* shRNA targeting the 3' UTR and ectopic *V5-PSMC2* expression. Data are presented as averages ± S.D.

The increased vulnerability of *PSMC2*^{Loss} lines correlated with both *PSMC2* copy loss and lower *PSMC2* mRNA expression (FDR p-value<0.05 for both). Expression and copy-number of *PSMC2* are also correlated in both the CCLE (r = 0.64) and TCGA Ovarian (r = 0.49) sample sets (**Fig. 14a**), indicating that cancer cells that have *PSMC2* copy loss tolerate reduced *PSMC2* expression.

To explore the effects of *PSMC2* loss on PSMC2 protein levels, we evaluated PSMC2 levels in IOSE cells and ten ovarian cancer cell lines, including five *PSMC2*^{Neutral} and five *PSMC2*^{Loss} lines. To minimize potential confounding of other genetic events affecting the 19S complex, we selected *PSMC2*^{Neutral} lines that had no copy-number gains of *PSMC2* and *PSMC2*^{Loss} lines that had copy loss of no more than one other 19S regulatory complex subunit (**Table 10**). All five *PSMC2*^{Loss} cell lines expressed lower levels of PSMC2 than any of the other cell lines (**Fig. 14b**). In contrast, the levels of eight 19S subunits including PSMC1 (Rpt2), PSMC4 (Rpt3), PSMC6 (Rpt4), PSMC3 (Rpt5), PSMC5 (Rpt6), PSMD2 (Rpn1), PSMD1 (Rpn2), PSMD4 (Rpn10), or the 20S subunits PSMB5 (β5) and PSMA1-6 (α subunits) failed to correlate with *PSMC2* copy-number (**Fig. 14b**). Since PSMC2 is essential for cell proliferation, we concluded that *PSMC2*^{Neutral} cells either require more PSMC2 or produce more than is necessary for cell survival. We therefore investigated how *PSMC2*^{Neutral} cells can tolerate greater suppression of PSMC2.

Specifically, we expressed a *PSMC2*-specific shRNA under the control of a doxycycline-regulated promoter in OVCAR8 (*PSMC2*^{Loss}) and A2780 (*PSMC2*^{Neutral}) cells. The addition of doxycycline led to the suppression of PMSC2 in both OVCAR8 and A2780 cells (**Fig. 14c**). Under these conditions, A2780 cells continue to proliferate

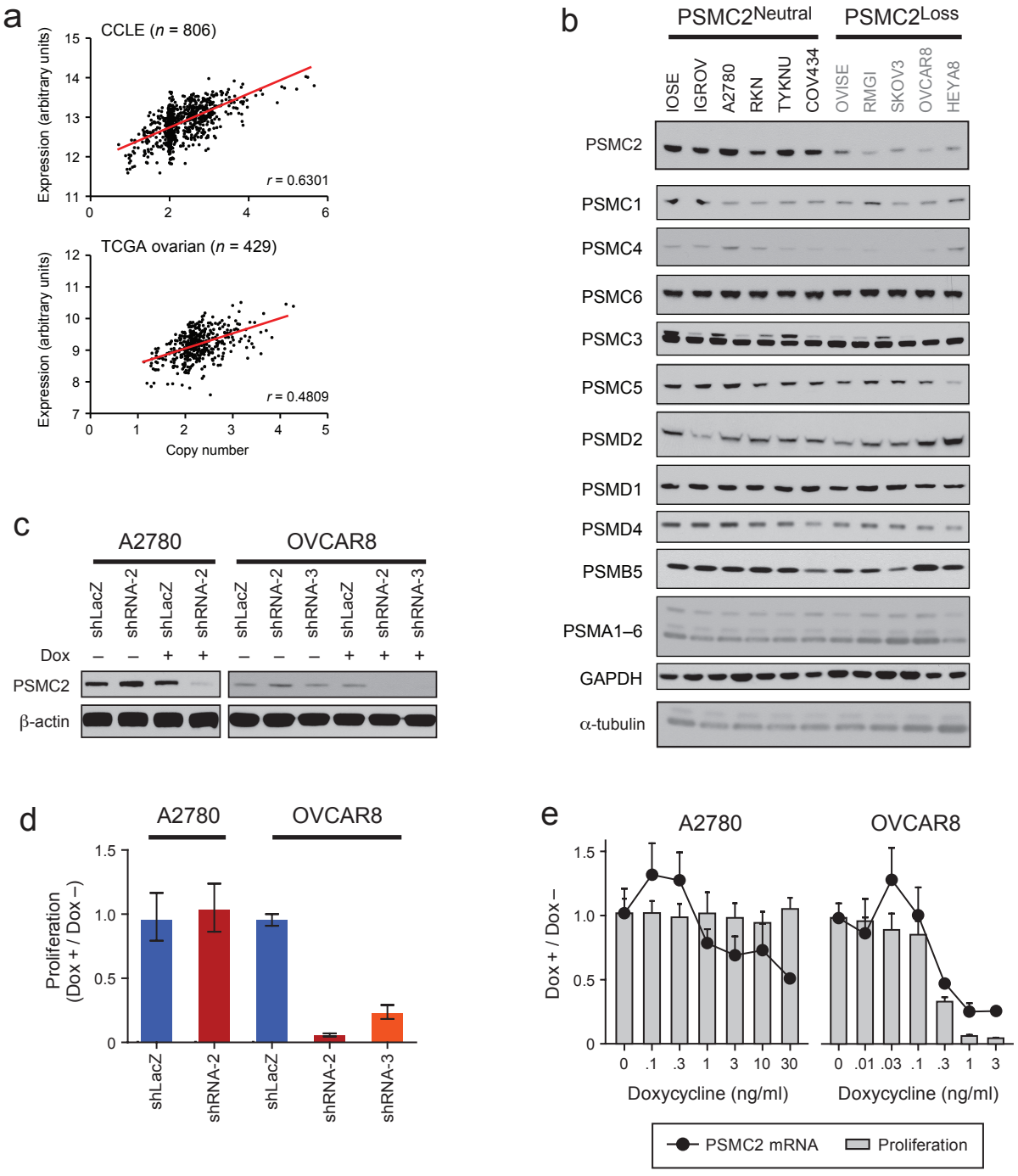

a
CCLE (n = 806)
Expression (arbitrary units)
r = 0.6301
TCGA ovarian (n = 429)
Expression (arbitrary units)
r = 0.4809
Copy number
b
PSMC2Neutral
PSMC2Loss
IOSE
IGROV
A2780
RKN
TYKNU
COV434
OVISE
RMGI
SKOV3
OVCAR8
HEYA8
PSMC2
PSMC1
PSMC4
PSMC6
PSMC3
PSMC5
PSMD2
PSMD1
PSMD4
PSMB5
PSMA1–6
GAPDH
α-tubulin
c
A2780
OVCAR8
shLacZ
shRNA-2
shRNA-3
Dox
PSMC2
β-actin
d
A2780
OVCAR8
Proliferation
(Dox + / Dox –)
shLacZ
shRNA-2
shRNA-3
e
A2780
OVCAR8
Dox + / Dox –
Doxycycline (ng/ml)
PSMC2 mRNA
Proliferation

Figure 14- Effects of PSMC2 on expression of proteasome components: (**a**) Scatterplots of PSMC2 copy number versus mRNA expression derived from cell lines in CCLE (Barretina et al., 2012) and primary ovarian tumors in TCGA. A linear regression line and the Pearson correlation coefficient (r) are also displayed. (**b**) Comparison between *PSMC2* copy number and protein expression of 10 other 26S proteasome components. (**c**) Immunoblot for PSMC2 in $PSMC2^{Loss}$ and $PSMC2^{Neutral}$ lines containing a Dox-inducible promotor for either shLacZ or shPSMC2. Cells were harvested 48 hours after addition of Dox, prior to qualitative observation of any phenotype. (**d**). Effects of Doxycycline-induced *PSMC2* suppression on proliferation. (**e**) Relationship between *PSMC2* mRNA expression and proliferation in $PSMC2^{Neutral}$ (left) and $PSMC2^{Loss}$ (right) cells. Data represents averages ± S.D.

Table10-Copy Number Status of 19S Proteasome Components Across Cell Lines

"1" Indicates Hemizygous Deletion; "0" indicates Copy Neutrality

gene	HEYA8	OVCAR8	OVISE	RMGI	SKOV3	A2780	IGROV1	OV90	TYKNU	RKN
PSMD4	0	1	0	0	0	0	0	0	0	0
PSMD14	1	1	0	0	0	0	0	0	0	0
PSMD1	1	1	0	0	0	0	0	0	0	1
PSMD6	0	0	1	0	0	0	0	1	0	0
PSMD2	0	1	0	0	0	0	0	0	0	0
PSMC2	1	1	1	1	1	0	0	0	0	0
PSMD5	0	0	0	0	0	0	0	0	0	0
PSMD13	1	1	0	0	0	0	0	1	0	0
PSMC3	1	1	0	0	0	0	0	0	0	0
PSMD9	0	0	0	0	0	0	0	0	0	0
PSMC6	0	0	0	0	0	0	0	0	0	0
PSMC1	0	1	0	0	0	0	0	0	0	0
PSMD7	1	1	0	0	0	0	0	0	0	0
PSMD11	1	1	0	0	0	0	0	0	0	0
PSMD3	0	1	0	0	0	0	0	0	0	0
PSMC3IP	0	0	0	0	0	0	0	0	0	0
PSMC5	0	0	0	0	0	0	0	0	0	0
PSMD12	0	0	0	0	0	0	0	0	0	0
PSMD8	0	1	0	0	0	0	0	0	0	0
PSMC4	0	1	0	0	0	0	0	0	0	0

whereas OVCAR8 cells arrest at G2/M and die by apoptosis, characteristic hallmarks of pharmacologic inhibition of the proteasome[347,348] (**Fig. 15**). To verify that A2780 cells tolerate a higher percentage of *PSMC2* suppression, we varied the degree of *PSMC2* mRNA suppression by changing the concentration of doxycycline from 0.01 ng/ml to 30 ng/ml, and found that a 50% decrease in *PSMC2* expression reduced the proliferation of OVCAR8 cells but had no effect on A2780 proliferation (**Fig. 14d-e**).

To determine the amount of PSMC2 required to maintain A2780 cell proliferation, we suppressed PSMC2 expression further by transfecting a pool of three *PSMC2* specific siRNAs at varying concentrations. Using quantitative RT-PCR and immunoblotting, we estimated that untreated OVCAR8 cells express approximately 50% of the *PSMC2* mRNA and protein found in A2780 cells (**Fig. 16a**) Furthermore, the proliferation of A2780 cells decreased only when *PSMC2* expression was suppressed by more than 60% (**Fig 16b**) and that both A2780 and OVCAR8 lose proliferative capacity at similar total levels of PSMC2 expression (**Fig. 16c**), suggesting that they have a comparable threshold requirement for PSMC2.

PSMC2^{Loss} cells exhibit only slight alterations in proteasome content and function

The tolerance of cells for loss of *PSMC2* copy-number and expression indicates that cells contain a reservoir of excess PSMC2 that is not required for proliferation. This reservoir may be maintained in an excess of fully assembled 26S proteasome or elsewhere in the cell. We analyzed proteasome assembly and content by performing Polyacrylamide Gel Electrophoresis (PAGE) on crude lysates under native (non-denaturing) conditions. Under these conditions, the 26S proteasome complex is stable

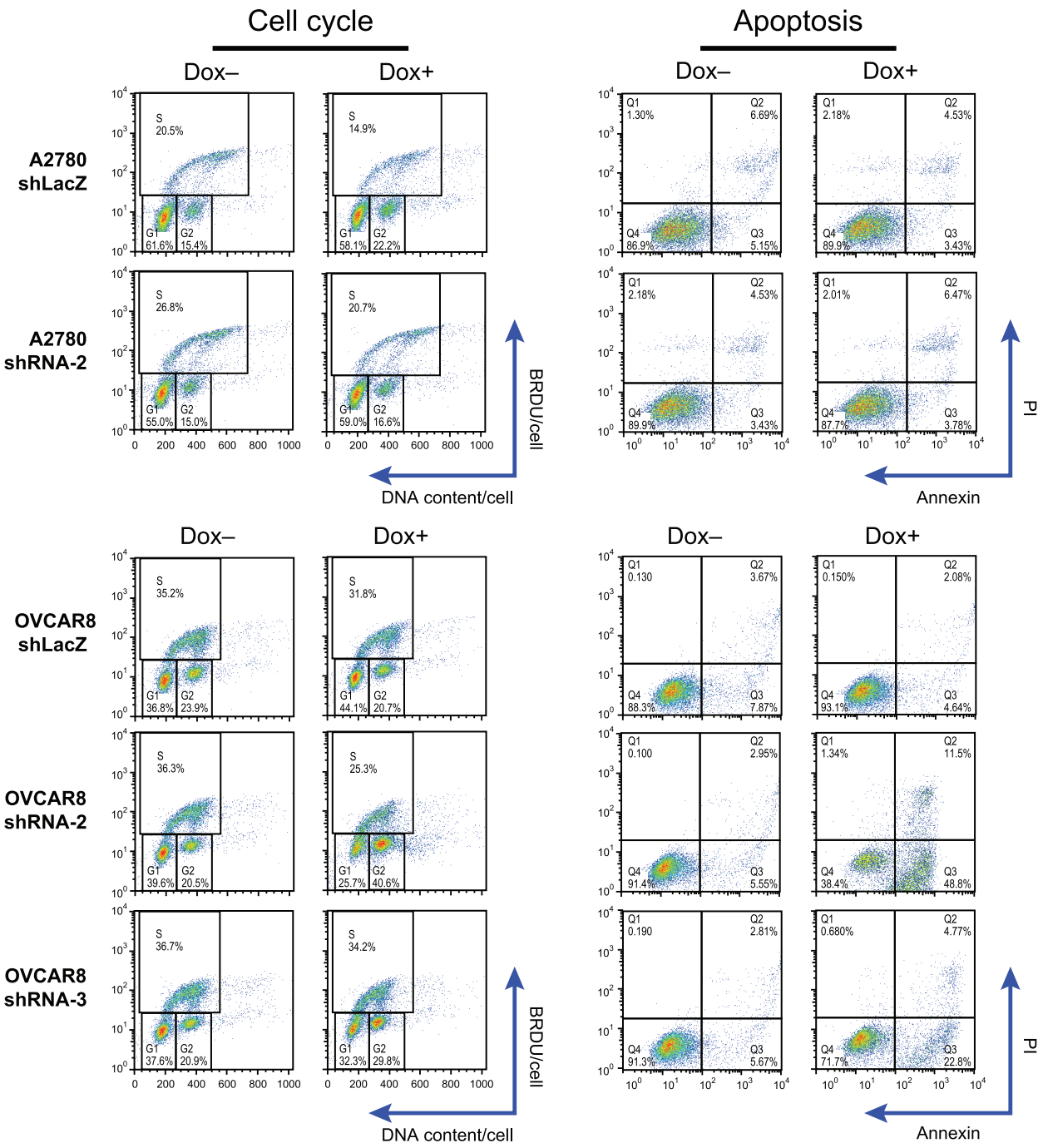


Figure 15- Effects of PSMC2 on cell cycle progression and apoptosis: Flow cytometry measurements of cell cycle progression (left) and apoptosis (right) in *PSMC2*^{Neutral} and *PSMC2*^{Loss} cells with and without suppression of *PSMC2*. No differences in cell cycle progression were observed, but *PSMC2*^{Loss} cells undergo increased Annexin V staining, indicating increased apoptosis, after *PSMC2* suppression.

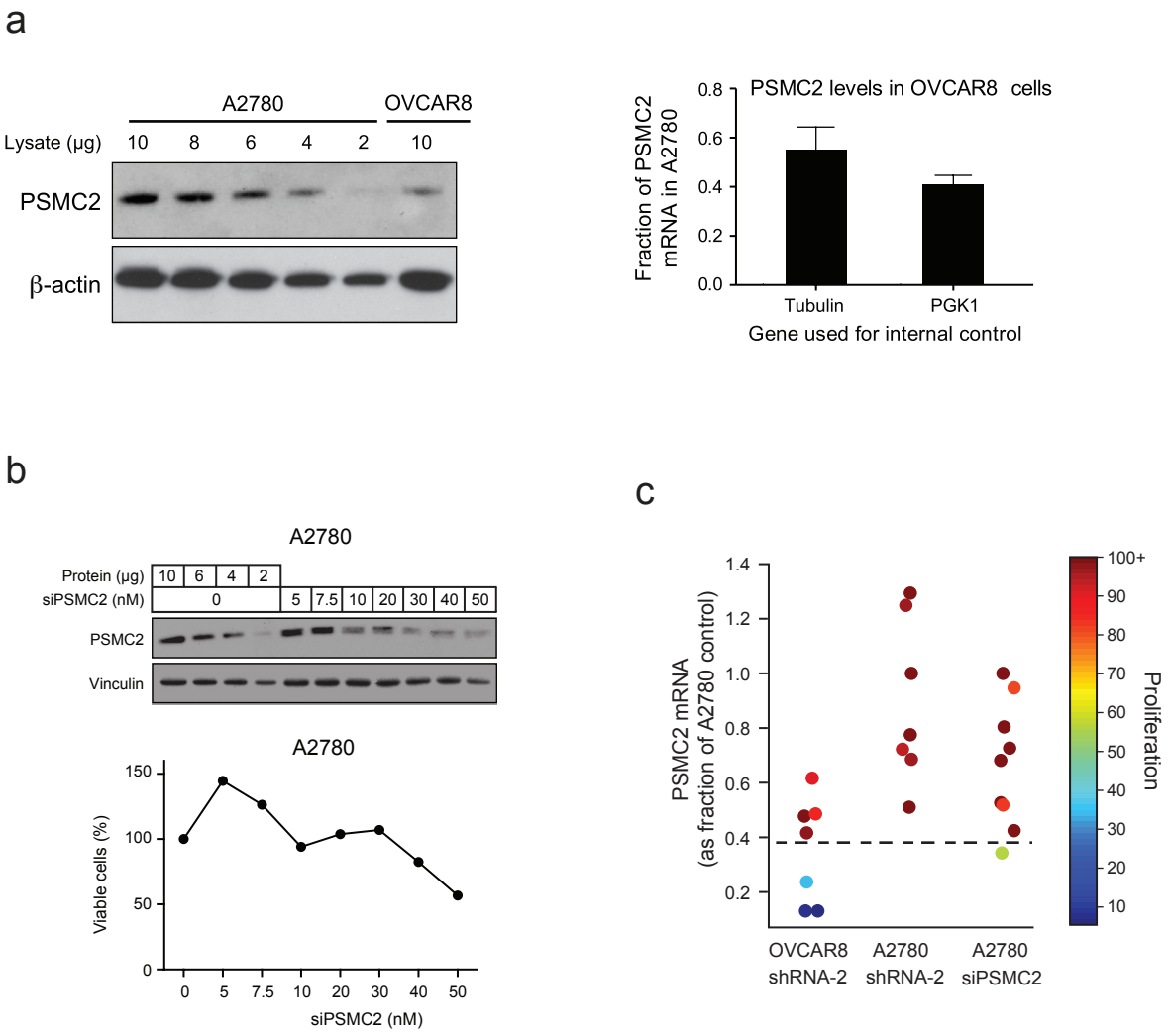


Figure 16- Effects of PSMC2 on cell cycle progression and apoptosis:
(**a**) Quantification of relative levels of PSMC2 protein. The immunoblot (left) indicates PSMC2 levels in OVCAR8 cell lysates relative to a dilution series using A2780 cell lysates. The graph on the right indicates fraction of *PSMC2* mRNA in OVCAR8 cells when compared to A2780, using two different genes to normalize qPCR product. (**b**) PSMC2 expression and relative proliferation in A2780 cells subjected to increasing levels of siRNA targeting *PSMC2*. (**c**) Schematic combining 26 data from Fig. 14d-e and Fig. 16b indicates that A2780 and OVCAR8 cells share a similar absolute threshold requirement for PSMC2 (dashed line).

and active and migrates in two distinct bands, distinguished by having either one or two 19S subunits incorporated in the formation of the $26S^{349}$. Using lysates collected from IOSE, two *PSMC2*^{Neutral}, and three *PSMC2*^{Loss} cancer cell lines (all with comparable proliferation rates), we detected $26S^1$, $26S^2$, and 20S proteasome complexes by immunoblotting for the core 20S subunits, PSMA1-6 (**Fig. 17a**).

We found that *PSMC2*^{Loss} lines express only slightly less 26S proteasome (most evident in $26S^2$), which is not comparable to the decrease in PSMC2 in these cells (**Fig. 17a**), and increased 20S proteasome. Similarly, comparable changes in PSMC2 expression in isogenic systems failed to substantially affect 26S proteasome content. Suppression of *PSMC2* levels by 50% in the Dox-shRNA-2 A2780 system led to an increase in the 20S complex but little to no change in $26S^1$ or $26S^2$ proteasome content relative to controls (**Fig. 17b**). Conversely, ectopic expression of PSMC2 in OVCAR8 cells led to a slight reduction in 20S levels and slight increases in $26S^1$ and $26S^2$ proteasome content (**Fig. 17b**). The levels of other 19S proteasome units remained unchanged (**Fig. 17c**).

Similarly, peptidase cleavage activity varied only slightly between *PSMC2*^{Neutral} and *PSMC2*^{Loss} lines. We observed the greatest differences in in-gel analyses of peptidase activity, which revealed less $26S^2$ proteasome peptidase cleavage and increased 20S peptidase activity in *PSMC2*^{Loss} cells (**Fig. 18a**). These changes were recapitulated by *PSMC2* suppression in A2780 cells and reversed by ectopic *PSMC2* expression in OVCAR8 cells (**Fig 18b**). The decrease in $26S^2$ activity in *PSMC2*^{Loss} relative to *PSMC2*^{Neutral} cells, however, was not associated with significant differences in peptidase cleavage when quantitatively assayed in whole cell lysates under conditions

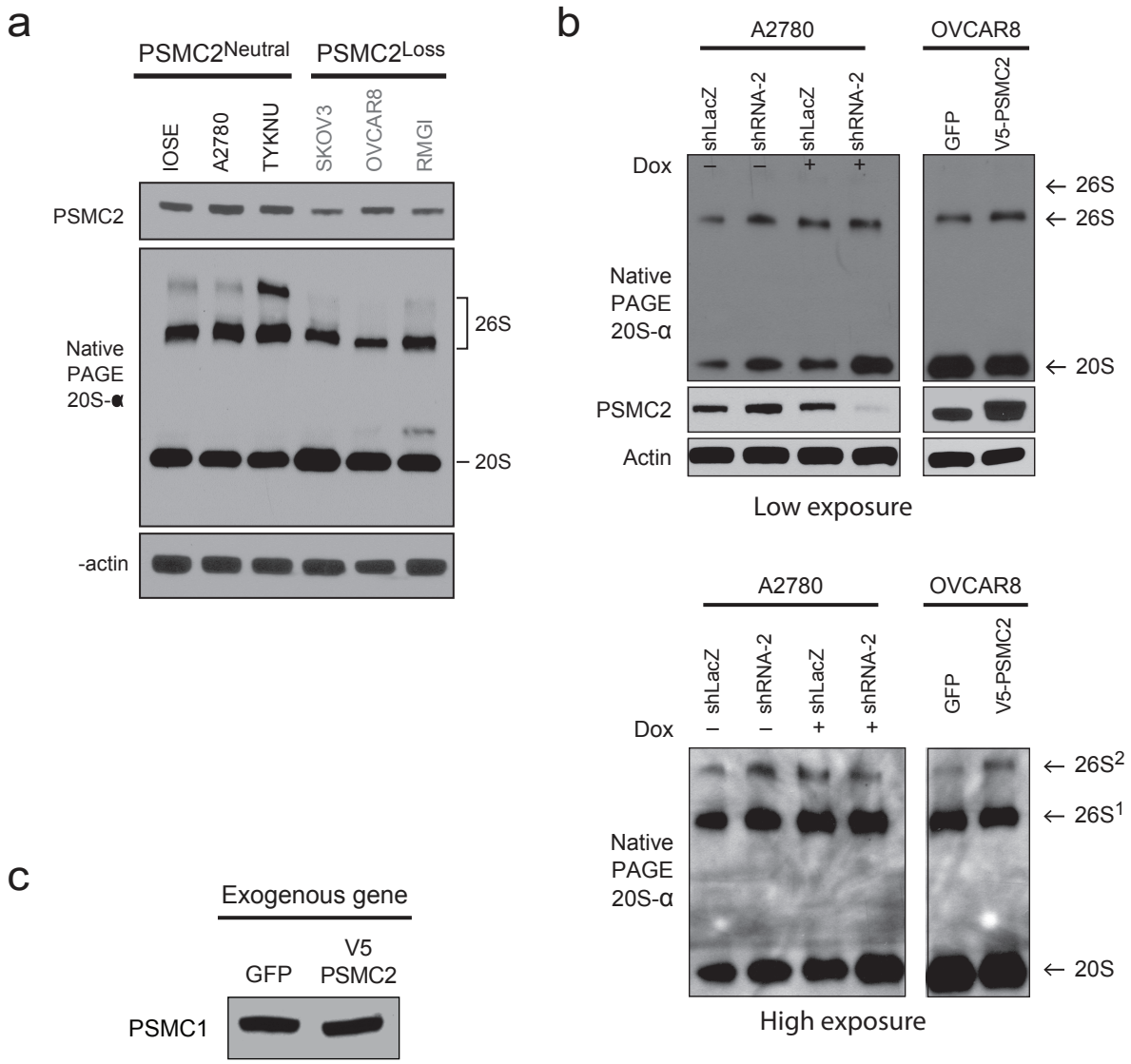


Figure 17: *PSMC2^{Loss}* cells lack a PSMC2 reservoir. (**a**) Total PSMC2 levels (top) and Native PAGE immunoblot for PSMA1-6 (middle) in *PSMC2^{Neutral}* and *PSMC2^{Loss}* cells. (**b**) Native PAGE immunoblot for PSMA1-6 in A2780 (left) and OVCAR8 (right) after inducible suppression or ectopic expression of *PSMC2*, respectively. Top and bottom show film from same immunoblot after different levels of exposure. (**c**) Immunoblot for PSMC3 and PSMC5 after ectopic expression of V5-PSMC2.

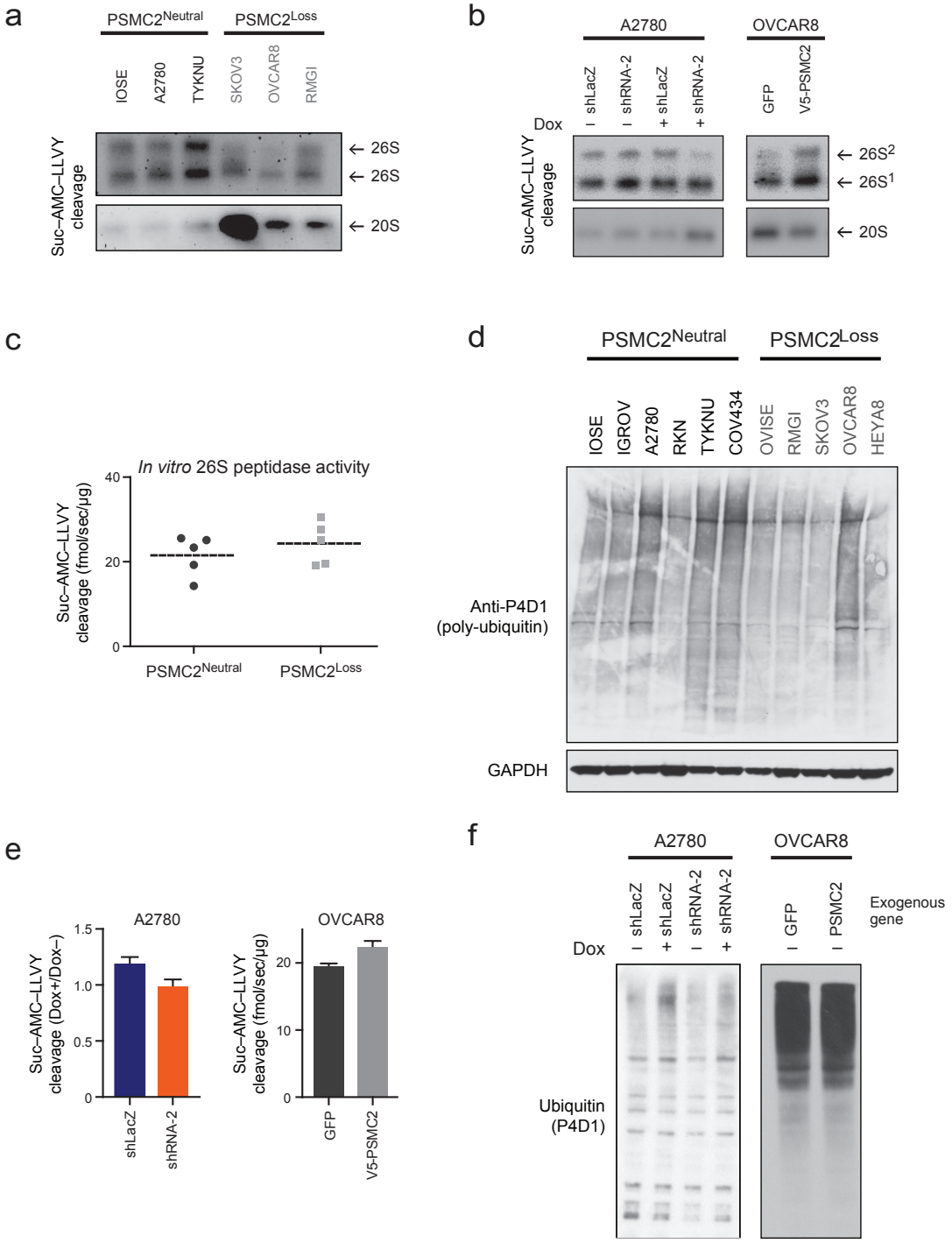


Figure 18: *PSMC2^{Loss}* cells lack a PSMC2 reservoir . Native PAGE 26S and 20S peptidase cleavage in *PSMC2^{Neutral}* and *PSMC2^{Loss}* cells (corresponding to 17a). Native PAGE 26S and 20S peptidase cleavage in isogenic systems (17b). (**c**) *In vitro* 26S proteasome activities in *PSMC2^{Neutral}* and *PSMC2^{Loss}* cells. Each point represents a cell line; dashed lines represent averages. We found no significant difference between the activity in these two populations. (**d**) Poly-ubiquitinated protein levels in *PSMC2^{Neutral}* and *PSMC2^{Loss}* cells**.** (**e**) *In vitro* 26S proteasome activities and (**f**) Poly-ubiquitinated protein levels for the isogenic systems used in (**b**).

(in the absence of SDS) in which free 20S proteasome does not contribute activity[350] (p=0.39) (**Fig. 18c**). In this assay, proteasome-specific peptidase activity is determined by bortezomib-inhibited cleavage. We found that 97% of activity was ablated by bortezomib, suggesting that other proteases did not contribute substantially to the measured activity. Lysates from *PSMC2*^{Neutral} and *PSMC2*^{Loss} lines grown under conventional non-stressed conditions also exhibited qualitatively similar total levels of poly-ubiquitin (**Fig. 18d**).

To test the acute effect of manipulating PSMC2 expression on peptidase activity, we measured peptidase activity in lysates of A2780 cells in which we suppressed *PSMC2* and of OVCAR8 cells engineered to recover *PSMC2* expression. Suppression of *PSMC2* by 50% in A2780 cells led to a 17% reduction in total 26S specific peptidase activity, associated with reduced $26S^2$ activity (**Fig. 18e**), neither of which corresponded to changes in total levels of poly ubiquitin (**Fig. 18f**). Conversely, ectopic *PSMC2* expression in OVCAR8 led to a 15% increase in peptidase activity, associated with increased $26S^2$ activity. The finding in both systems that modulating PSMC2 levels by up to 50% resulted in only a 17% alteration in 26S activity suggested that PSMC2 content was not the limiting component to 26S formation in *PSMC2*^{Neutral} cells.

We found no increased sensitivity to bortezomib in *PSMC2*^{Loss} cells and no significant correlation between the concentration of bortezomib that inhibits proliferation by 50% (IC_{50}) and decreased expression of any of the 47 26S proteasome components, across 133 cell lines previously tested[351] (**Table 11**). Suppression of *PSMC2* in Dox-shRNA-2 A2780 cells or ectopic *PSMC2* expression in OVCAR8 cells also did not substantially affect the bortezomib IC_{50} (**Fig. 19a-b**). These observations are consistent

Table 11-Analysis to correlate Bortezomib IC_{50} with expression of proteasome components

Left tail = genes where high expression correlates with low bortezomib IC_{50};

Right tail = genes where low expression correlates with low bortezomib IC50

Proteasome Component	R sided p value	Left sided p value	Right sided FDR	Left sided FDR
PSMG4	0.9983	0.0017	0.9983	0.0799
PSMD5	0.9788	0.0212	0.9983	0.488
PSMC2	0.9595	0.0405	0.9983	0.488
PSMB2	0.9521	0.0479	0.9983	0.488
PSMD2	0.9481	0.0519	0.9983	0.488
PSMB7	0.9371	0.0629	0.9983	0.488
PSMD12	0.9243	0.0757	0.9983	0.488
PSMD14	0.9042	0.0958	0.9983	0.488
PSMG1	0.904	0.096	0.9983	0.488
PSMD1	0.8962	0.1038	0.9983	0.488
PSME2	0.8712	0.1288	0.9983	0.5497
PSMA7	0.8248	0.1752	0.9983	0.5497
PSMC4	0.8241	0.1759	0.9983	0.5497
PSMB6	0.8225	0.1775	0.9983	0.5497
PSMB8	0.8134	0.1866	0.9983	0.5497
PSMD6	0.8129	0.1871	0.9983	0.5497
PSMA2	0.7981	0.2019	0.9983	0.5581
PSMD10	0.7537	0.2463	0.9983	0.634
PSMB9	0.7437	0.2563	0.9983	0.634
PSMC5	0.7105	0.2895	0.9983	0.6636
PSMB5	0.6971	0.3029	0.9983	0.6636
PSMA3	0.6719	0.3281	0.9983	0.6636
PSMA5	0.6567	0.3433	0.9983	0.6636
PSME1	0.6414	0.3586	0.9983	0.6636
PSMG2	0.6158	0.3842	0.9983	0.6636
PSMD11	0.6133	0.3867	0.9983	0.6636
PSME3	0.6115	0.3885	0.9983	0.6636
PSMC6	0.6022	0.3978	0.9983	0.6636
PSMB3	0.5803	0.4197	0.9983	0.6636
PSMG3	0.5671	0.4329	0.9983	0.6636
PSMA1	0.5496	0.4504	0.9983	0.6636
PSMC3IP	0.5482	0.4518	0.9983	0.6636
PSMD9	0.5338	0.4662	0.9983	0.6639
PSMA8	0.5054	0.4946	0.9983	0.6838
PSMB1	0.4777	0.5223	0.9983	0.7014
PSMB10	0.4346	0.5654	0.9983	0.7303
PSMA4	0.4171	0.5829	0.9983	0.7303
PSMD7	0.3949	0.6051	0.9983	0.7303
PSMD8	0.394	0.606	0.9983	0.7303
PSMB4	0.297	0.703	0.9983	0.826
PSMD3	0.2699	0.7301	0.9983	0.837
PSMD13	0.1462	0.8538	0.9983	0.9526
PSMD4	0.1232	0.8768	0.9983	0.9526
PSMF1	0.1082	0.8918	0.9983	0.9526
PSME4	0.0686	0.9314	0.9983	0.9646
PSMC3	0.0451	0.9549	0.9983	0.9646
PSMA6	0.0354	0.9646	0.9983	0.9646

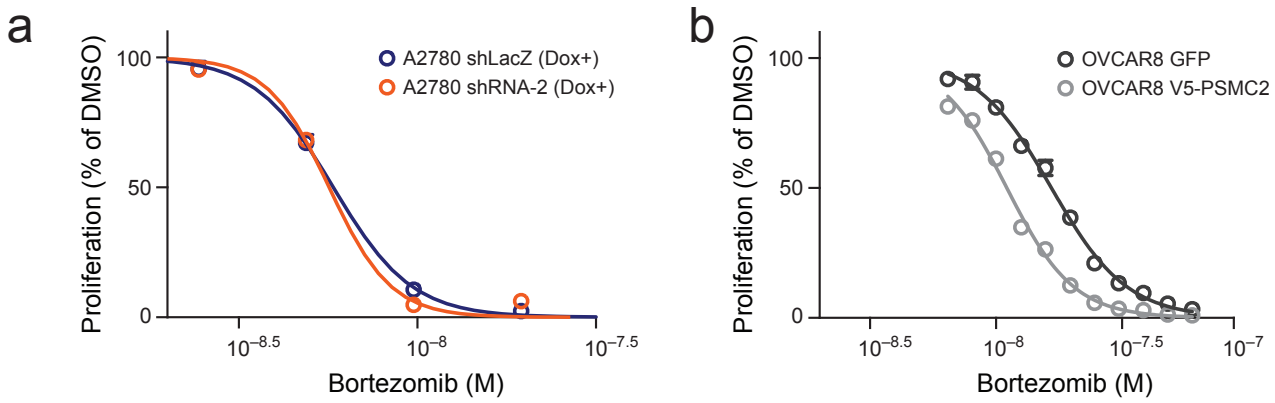


Figure 19- PSMC2^{Loss} cells and sensitivity to bortezomib: (**a-b**) Dose response curve for bortezomib in (**a**) A2780 cells with and without *PSMC2* suppression and (**b**) OVCAR8 with and without ectopic V5-*PSMC2* expression.

with our prior observation that 26S proteasome function is not substantially compromised in *PSMC2*^{Loss} cells.

PSMC2^{Neutral} cells have a reservoir of PSMC2 that buffers 26S proteasome levels against PSMC2 loss

The finding that *PSMC2*^{Neutral} cells have near equal 26S proteasome content to *PSMC2*^{Loss} cells even though they express higher levels of PSMC2 suggests that *PSMC2*^{Neutral} cells contain a separate reservoir of PSMC2 that is preferentially lost when levels are reduced. To identify this reservoir, we combined native PAGE with immunoblotting for PSMC2 across a panel of cell lines (**Fig. 20a**). Of the multiple reactive bands identified, even after a long exposure, only one band ("Complex^{PSMC2}") was present in all of the *PSMC2*^{Neutral} but none of the *PSMC2*^{Loss} lines. Using isogenic systems, we also found that *PSMC2* suppression in Dox-shRNA-2 A2780 cells led to reduced levels of Complex^{PSMC2}, whereas ectopic PSMC2 expression in OVCAR8 cells led to its reappearance (**Fig. 20b**). These results suggest that Complex^{PSMC2} is a specific PSMC2 reservoir.

We hypothesized that Complex^{PSMC2} serves as a "buffer" in *PSMC2*^{Neutral} cells, enabling such cells to maintain 26S proteasome levels and function in the face of reduced PSMC2 expression. In this case, *PSMC2* suppression should deplete Complex^{PSMC2} before reducing 26S proteasome levels. To quantify the consequences of reducing PSMC2 on Complex^{PSMC2} and 26S proteasome levels, we compared dilutions of lysates from Dox shRNA-2 A2780 cells propagated in the absence of doxycycline to lysate collected from these cells cultured in doxycycline (**Fig. 20c**). In cells in which

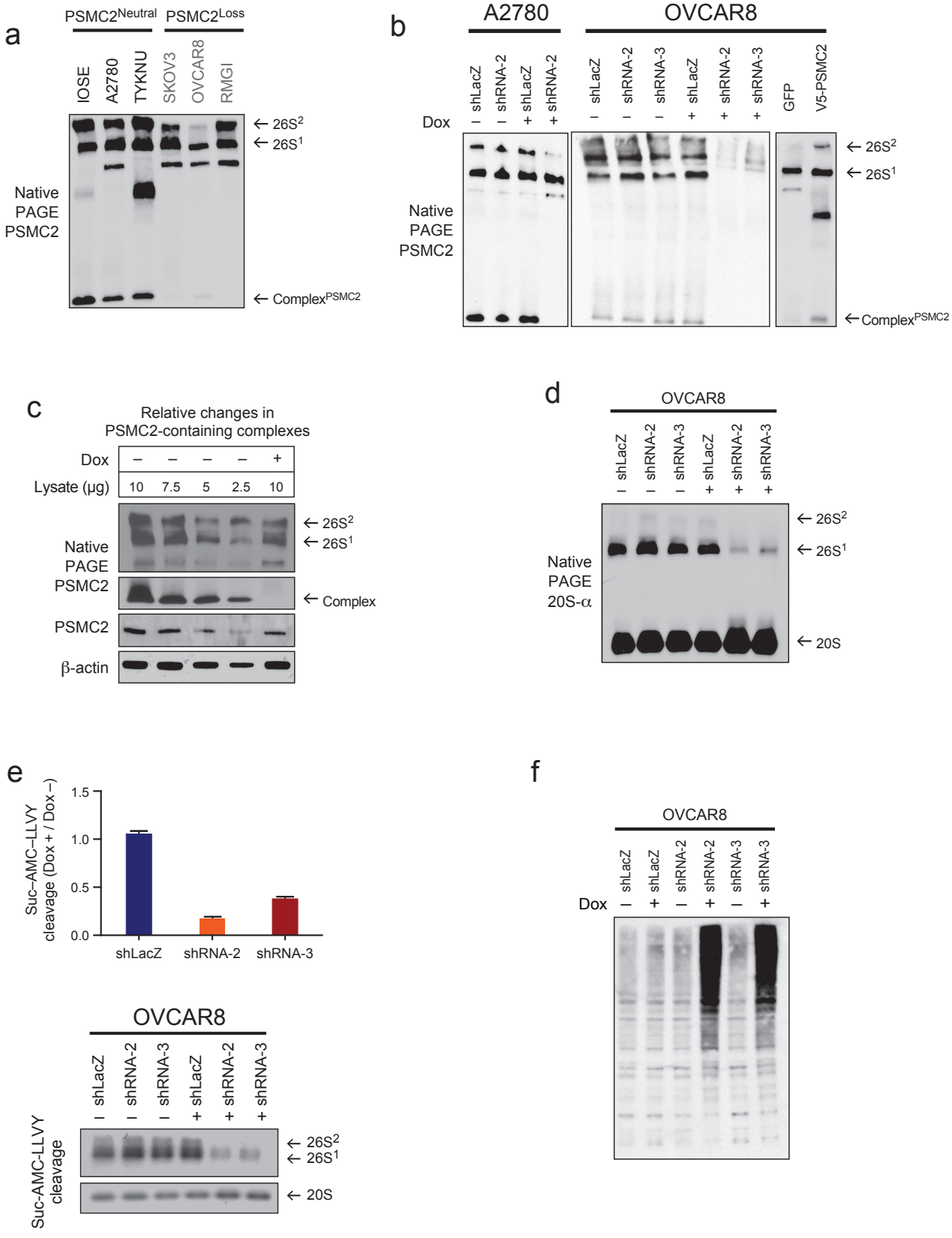

a
PSMC2Neutral
PSMC2Loss
IOSE
A2780
TYKNU
SKOV3
OVCAR8
RMGI
Native PAGE PSMC2
← 26S2
← 26S1
← ComplexPSMC2
b
A2780
OVCAR8
Dox
– shLacZ
– shRNA-2
+ shLacZ
+ shRNA-2
– shLacZ
– shRNA-2
– shRNA-3
+ shLacZ
+ shRNA-2
+ shRNA-3
GFP
V5-PSMC2
Native PAGE PSMC2
← 26S2
← 26S1
← ComplexPSMC2
c
Relative changes in PSMC2-containing complexes
Dox – – – – +
Lysate (μg) 10 7.5 5 2.5 10
Native PAGE PSMC2
PSMC2
β-actin
← 26S2
← 26S1
← Complex
d
OVCAR8
– shLacZ
– shRNA-2
– shRNA-3
+ shLacZ
+ shRNA-2
+ shRNA-3
Native PAGE 20S-α
← 26S2
← 26S1
← 20S
e
Suc–AMC–LLVY cleavage (Dox + / Dox –)
1.5
1.0
0.5
0.0
shLacZ
shRNA-2
shRNA-3
OVCAR8
– shLacZ
– shRNA-2
– shRNA-3
+ shLacZ
+ shRNA-2
+ shRNA-3
Suc-AMC-LLVY cleavage
← 26S2
← 26S1
← 20S
f
OVCAR8
Dox
– shLacZ
+ shLacZ
– shRNA-2
+ shRNA-2
– shRNA-3
+ shRNA-3

Figure 20. Complex^{PSMC2} buffers *PSMC2^{Neutral}* cells against *PSMC2* suppression: (**a**) Native PAGE immunoblot for PSMC2 across a panel of *PSMC2^{Neutral}* and *PSMC2^{Loss}* cells. (**b**) Native PAGE immunoblot for PSMC2 in A2780 after inducible expression (left), and in OVCAR8 cells after inducible suppression or ectopic expression, of PSMC2. (c) Quantification of 26S proteasome and Complex^{PSMC2} levels after *PSMC2* suppression in Dox-shRNA-2 A2780 cells by Native PAGE (top) and total PSMC2 levels (bottom). (**d-f)** OVCAR8 cells with and without *PSMC2* suppression analyzed by Native PAGE immunoblots for (**d**) PSMA1-6 and (**e**) peptidase cleavage measured either in whole-cell and native gel formats (see **Fig. 18a-c**), and (**f**) total poly-ubiquitin levels

PSMC2 was suppressed, the relative loss of Complex^{PSMC2} exceeded the decrease in 26S proteasome content. These observations indicate that Complex^{PSMC2} was preferentially lost in A2780 cells after *PSMC2* suppression. In contrast, *PSMC2* suppression in OVCAR8 cells, which lack Complex^{PSMC2}, led to near-complete ablation of 26S proteasome levels (**Fig. 20b**) and peptidase activity (**Fig. 20e**) and to a qualitative increase in the amount of poly-ubiquitin (**Fig. 20f**).

To analyze the components of Complex^{PSMC2}, we fractionated lysates from IOSE cells expressing either V5-GFP or V5-PSMC2 (**Fig. 21a**) using a glycerol gradient (**Fig. 21b**), and isolated V5-immune complexes containing either Complex^{PSMC2} or 26S proteasome. Complex^{PSMC2} immune complexes (collected in fractions 2-4) contained PSMC2, PSMC1 (Rpt2), PSMD2 (Rpn1), and PSMD5 (S5B) (**Fig. 21c**), subunits of one of three complexes known to compose the base of the 19S proteasome [330-332,352,353]. Complex^{PSMC2} did not contain subunits of the other two complexes, PSMC3 (Rpt5), PSMC4 (Rpt3), PSMC5 (Rpt6), PSMC6 (Rpt6), or members of the 20S proteasome, PSMB5 (β5), or PSMA1-6 (α subunits) (**Fig. 21c**). All of these proteins except PSMD5 were detected in immune complexes containing the 26S complex (from fractions 7-9). These observations indicate that the PSMC2 reservoir is a subcomplex of the 26S proteasome.

The reduction of PSMC2 levels in *PSMC2^{Loss}* cells inhibits orthotopic tumor growth

To explore the therapeutic potential of *PSMC2* suppression *in vivo*, we tested the consequences of suppressing PSMC2 in ovarian xenografts. Specifically, we used a

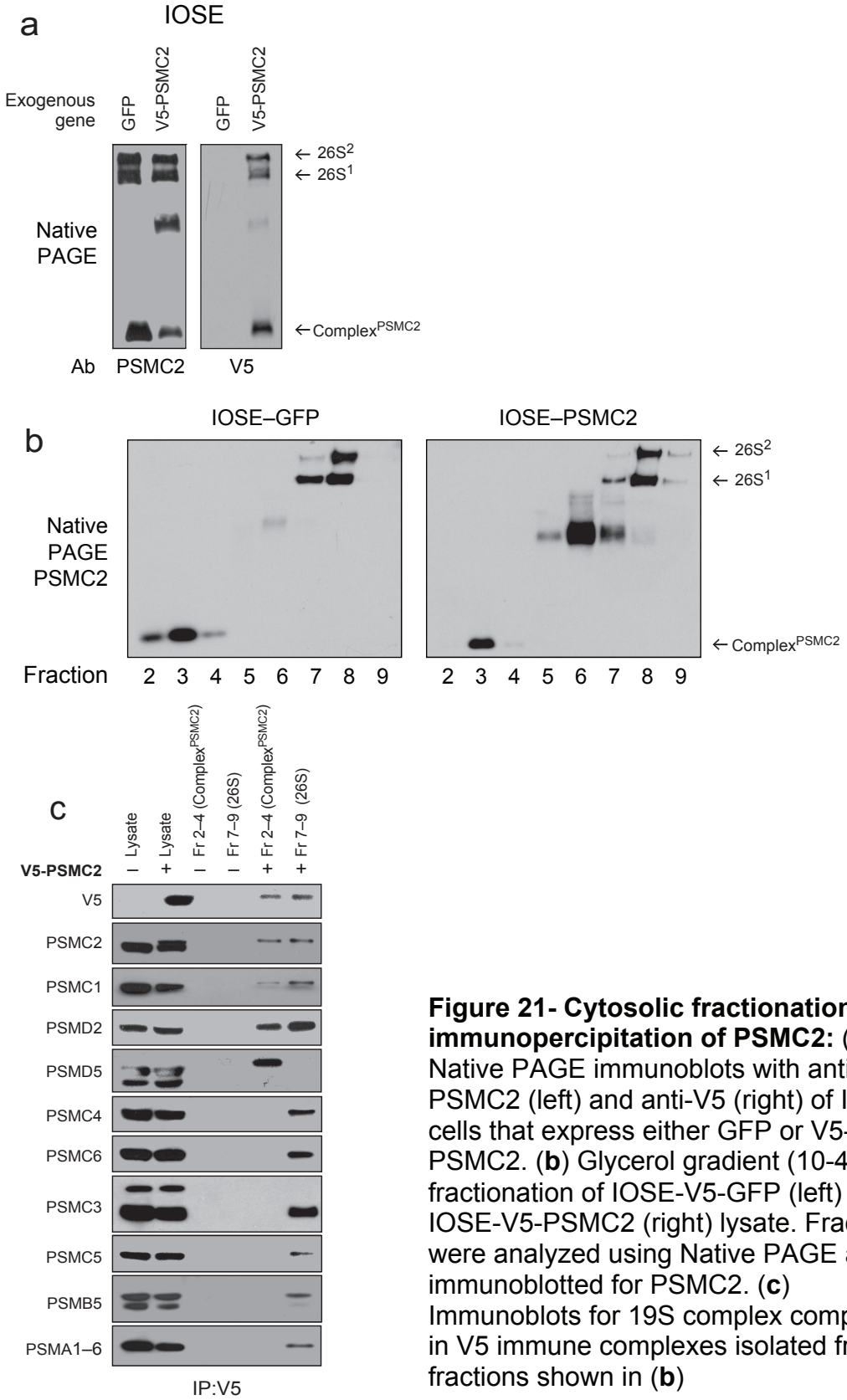


Figure 21- Cytosolic fractionation and immunopercipitation of PSMC2: (**a**) Native PAGE immunoblots with anti-PSMC2 (left) and anti-V5 (right) of IOSE cells that express either GFP or V5-PSMC2. (**b**) Glycerol gradient (10-40%) fractionation of IOSE-V5-GFP (left) and IOSE-V5-PSMC2 (right) lysate. Fractions were analyzed using Native PAGE and immunoblotted for PSMC2. (**c**) Immunoblots for 19S complex components in V5 immune complexes isolated from fractions shown in (**b**)

tumor-targeted nanoparticle delivery system that delivers siRNA into the cytosol of cells within the tumor parenchyma[345]. We generated tumor-penetrating nanocomplexes (TPNs) consisting of *PSMC2*-specific siRNA non-covalently bound to tandem peptides bearing an N-terminal cell-penetrating domain, Transportan (TP), and a C-terminal tumor-specific domain, LyP-1 (CGNKRTRGC), which binds to its cognate receptor p32 (**Fig. 22a**).

We first assessed the compatibility of cell lines with TPN-targeted siRNA delivery. OVCAR8 and A2780 cells exhibited high cell surface levels of expression of p32, whereas IOSE cells exhibited low expression (**Fig 22b**). In consonance with these observations, flow cytometry to quantify cytosolic delivery of fluorescently labeled siRNAs indicated substantial accumulation of siRNA in both OVCAR8 and A2780 cells (**Fig. 22c**). A monoclonal antibody directed against p32 (mAb 60.11) substantially reduced nanocomplex uptake, whereas a control antibody had no effect on uptake. These results indicate that surface p32 expression correlates with enhanced uptake of TPNs and that TPN-mediated siRNA delivery is p32 receptor-specific.

We next used these TPNs to confirm the vulnerability of $PSMC2^{Loss}$ cells to *PSMC2* suppression both *in vitro* and *in vivo*. We treated OVCAR8 and A2780 cells *in vitro* with TPNs carrying siRNAs targeting non-overlapping exons of *PSMC2*. In both cell types, we observed a reduction of PSMC2 protein relative to cells treated with TPNs carrying *GFP* siRNA (**Fig. 22d**). This reduction was associated with a corresponding decrease in proliferation only in the OVCAR8 cells (**Fig. 22e**). We then used these TPNs to treat mice harboring orthotopic OVCAR8 or A2780 tumors expressing firefly luciferase. We injected TPNs carrying *PSMC2*-siRNA (1 mg siRNA/kg body weight for

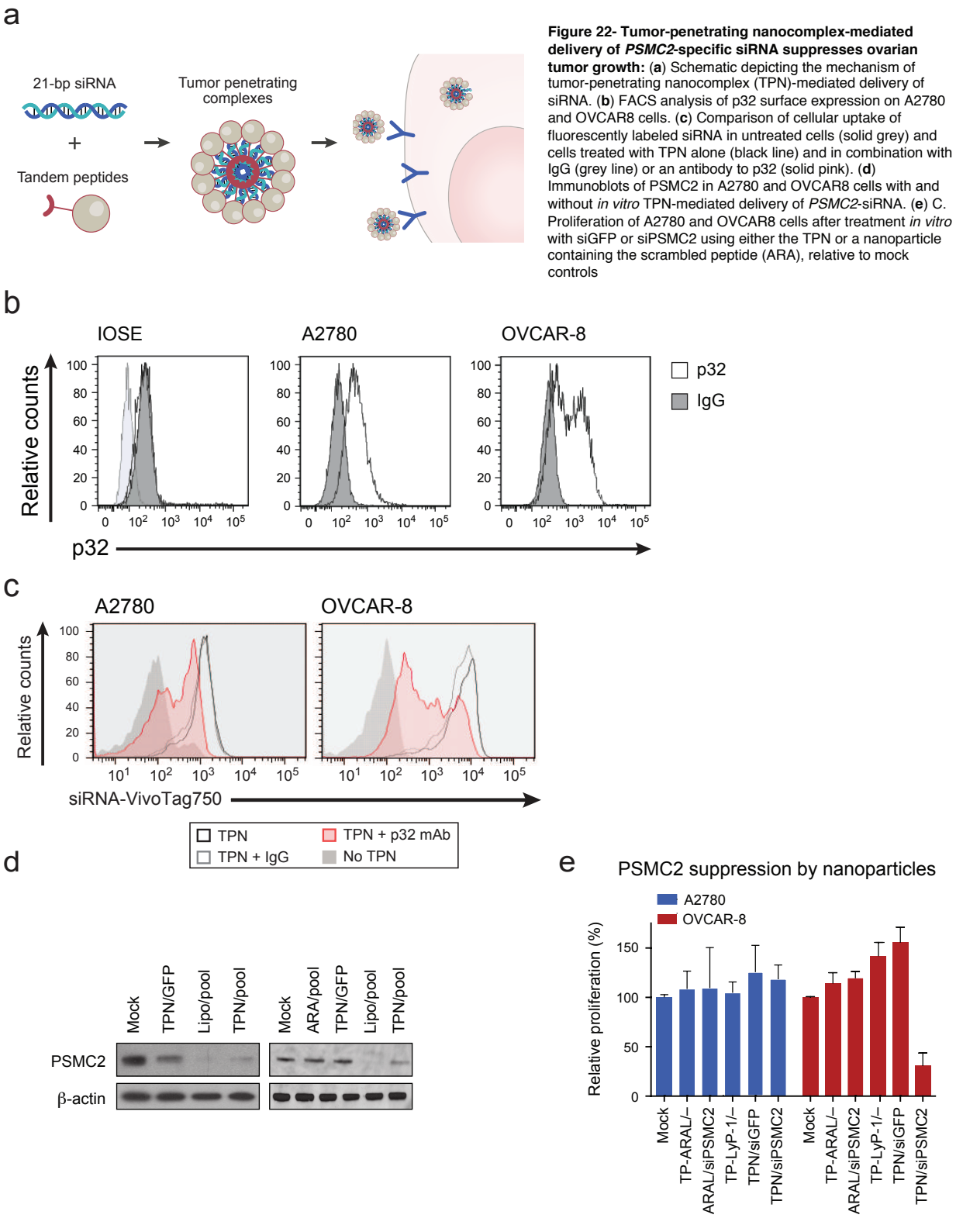


Figure 22- Tumor-penetrating nanocomplex-mediated delivery of *PSMC2*-specific siRNA suppresses ovarian tumor growth: (**a**) Schematic depicting the mechanism of tumor-penetrating nanocomplex (TPN)-mediated delivery of siRNA. (**b**) FACS analysis of p32 surface expression on A2780 and OVCAR8 cells. (**c**) Comparison of cellular uptake of fluorescently labeled siRNA in untreated cells (solid grey) and cells treated with TPN alone (black line) and in combination with IgG (grey line) or an antibody to p32 (solid pink). (**d**) Immunoblots of PSMC2 in A2780 and OVCAR8 cells with and without *in vitro* TPN-mediated delivery of *PSMC2*-siRNA. (**e**) C. Proliferation of A2780 and OVCAR8 cells after treatment *in vitro* with siGFP or siPSMC2 using either the TPN or a nanoparticle containing the scrambled peptide (ARA), relative to mock controls

14 days) intraperitoneally every three days and monitored tumor burden non-invasively by imaging bioluminescence. We observed a reduction in tumor burden (by >75% relative to tumors treated with *siGFP*) only in OVCAR8 tumors (**Fig. 23a**). A2780 and any remaining OVCAR8 tumors treated with TPN/siPSMC2 exhibited lower levels of PSMC2 but not two other members of Complex^{PSMC2}, PSMC1 and PSMD5 (**Fig 23b**).

However, TPN/si*PSMC2* nanoparticles failed to decrease tumor burden of *PSMC2*^{Loss} cells in which we reconstituted PSMC2 expression *in vivo* using orthotopic tumor xenografts derived from OVCAR8 cells expressing V5-PSMC2 (**Fig. 23c**). This finding confirmed that the effects of TPN/siPSMC2 on tumor growth were the consequence of reduced PSMC2 expression.

Conversely, TPN/si*PSMC2* nanoparticles reduced tumor growth and significantly improved survival in *PSMC2*^{Neutral} cells expressing *PSMC2*-specific shRNAs (**Fig . 23e**). We measured the effects of TPN/si*PSMC2* nanoparticles relative to TPN/si*GFP* or PBS in mice with xenografts of A2780 cells engineered to express inducible *PSMC2* shRNA. Among mice treated with doxycycline and TPN/si*PSMC2*, overall survival was 40 days and 40% survived more than 42 days, whereas all animals in the TPN/si*GFP* and PBS cohorts succumbed to tumors within 19 days (p=0.0013) (**Fig. 23e**). These findings demonstrated the therapeutic efficacy of *PSMC2* suppression *in vivo*, and support the notion that *PSMC2*^{Loss} cells are sensitive to suppression of PSMC2 due to decreased basal levels of *PSMC2* mRNA.

Discussion

From a therapeutic standpoint, increasing the number of targetable cancer dependencies has the potential to increase the number of treatable patients, and focusing on driver alterations limits us to targeted therapies involving these events.

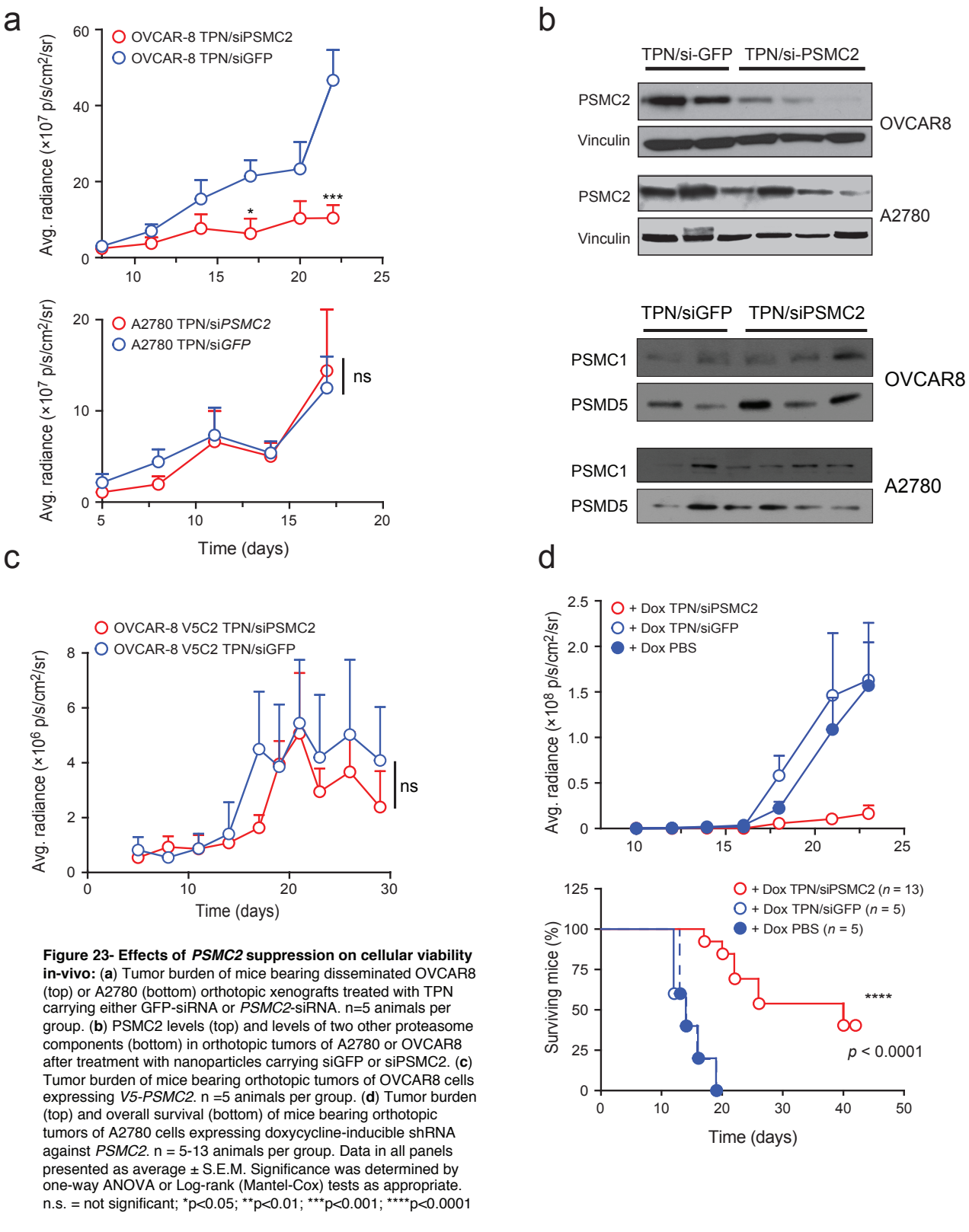


Figure 23- Effects of *PSMC2* suppression on cellular viability in-vivo: (**a**) Tumor burden of mice bearing disseminated OVCAR8 (top) or A2780 (bottom) orthotopic xenografts treated with TPN carrying either GFP-siRNA or *PSMC2*-siRNA. n=5 animals per group. (**b**) PSMC2 levels (top) and levels of two other proteasome components (bottom) in orthotopic tumors of A2780 or OVCAR8 after treatment with nanoparticles carrying siGFP or siPSMC2. (**c**) Tumor burden of mice bearing orthotopic tumors of OVCAR8 cells expressing *V5-PSMC2*. n =5 animals per group. (**d**) Tumor burden (top) and overall survival (bottom) of mice bearing orthotopic tumors of A2780 cells expressing doxycycline-inducible shRNA against *PSMC2*. n = 5-13 animals per group. Data in all panels presented as average ± S.E.M. Significance was determined by one-way ANOVA or Log-rank (Mantel-Cox) tests as appropriate. n.s. = not significant; *p<0.05; **p<0.01; ***p<0.001; ****p<0.0001

Additionally, many driver alterations belong to categories historically refractory to reversal with small molecule modulators[220,354-356], which may leave many patients with no targetable driver alterations. In such cases, having other targets of opportunity will be vital.

Even when targeting driver alterations, treatment refractory recurrence has been almost inevitable[357-359]. Rather than evidence against chasing passenger vulnerabilities, the repeated failure to find persistent dependence, even in the most likely targets, should further motivate us to expand our search for effective therapies to non-traditional targets. It is increasingly obvious that targeted therapies, acting in isolation, are unlikely to lead to sustained remission, and only by targeting multiple pathways simultaneously can we avoid recurrence caused by emergence of resistant subclones. Non-driver dependencies, if not effective in isolation can be tool in such combination therapy approaches.

PSMC2 as a CYCLOPS gene

By integrating data derived from the genomic characterization of human tumors with systematic interrogation of essential genes in cancer cell lines, we have identified a distinct class of cancer specific vulnerabilities associated with partial copy-number loss of essential genes. Hemizygous loss of *PSMC2* in particular, and of CYCLOPS genes in general, renders cells highly dependent on the remaining allele. Although *PSMC2* is frequently involved in partial copy-number loss, we did not observe homozygous deletion, consistent with the notion that *PSMC2* is an essential gene. Partial copy-number loss, in contrast, resulted in no measurable impact on either proteasome function or cell proliferation. Specifically, ectopic expression of PSMC2 failed to

enhance cell proliferation, and copy-number loss of *PSMC2* was not associated with decreased proteasome activity. Taken together these observations define an approach to identify a new class of context specific cancer dependencies.

26S proteasome components are not in stoichiometric equilibrium within cells, and the limiting components may differ between cancer and normal cells. For example, cells often express free 20S complex, but not 19S, suggesting that 26S proteasome levels are limited by the levels of 19S regulatory complex (**Fig. 17a**). There may be a similar imbalance between the modules that make up the base of the 19S complex. We found that the module containing PSMC2 (Rpt1), PSMC1 (Rpt2), PSMD2 (Rpn1), and PSMD5 (S5B) was in excess in many cancer cell lines, yet became limiting to 19S formation in $PSMC2^{Loss}$ cells, unveiling a new sensitivity. PSMC2 levels are influenced by its sub-complex partners[360], so interfering with the formation of $Complex^{PSMC2}$ or of its incorporation into the 19S proteasome may be a specific approach to reduce PSMC2 levels and proliferation in $PSMC2^{Loss}$ cells. Indeed, when we suppressed PSMC2 *in vivo* using tumor penetrating nanoparticles, we were able to obtain >75% reductions in tumor burden and a doubling of overall survival.

Since the proteasome is essential in all cells, one concern in targeting CYCLOPS gene targets is whether this strategy would induce substantial toxicity to non-cancer cells. However, proteasome inhibition has been well tolerated in humans, with bortezomib treatment resulting in 70% reduction of proteasome activity as measured by LLVY peptide cleavage, with acceptable side effect profiles[361]. In comparison, proliferation of $PSMC2^{Loss}$ cells is reduced at levels of PSMC2 suppression that result in only a 15% reduction of peptide cleavage in $PSMC2^{Neutral}$ cells.

Our findings suggest that one consequence of genomic instability is an alteration in the stoichiometry of components of macromolecular machines including the proteasome, ribosome, and spliceosome. These observations suggest that many of these imbalances may present potential therapeutic targets in individual components or precursor complexes, and that these components, rather than the fully assembled machines, will require specific inhibition or disruption.

Perspectives and Future Directions

SCNA Determination and future significance analyses

Discrete, allelic SCNA data provides a much richer source of information for future significance analyses. In the short term, copy neutral loss of heterozygosity (cnLOH) events can be paired with exomic sequencing across many of the TCGA samples to better identify potential recessive tumor suppressor genes.

We are also developing significance analyses specifically designed to take advantage of the data from this platform and model both positive and negative selection of SCNAs acquired through cancer development. In current analyses, the basic assumption is that positive selection through strong driver genes is the primary force shaping the somatic alterations we observe. With this assumption, we can assume a uniform background of expected somatic alterations, and deviations from this are seen as likely driver events in cancer development. However, many recent studies have implied that much of what we observe is shaped by either strong negative selection, such as inactivation of essential genes, or from accumulation of modest selective pressures[362].

In addition, our group and others have shown that the location of a gene within the chromosome may significantly affect the background rate of alterations at that locus[19,21]. With this in mind, one can imagine that the types of genes neighboring driver gene may have just as much influence on the SCNA profile at that locus as the driver gene itself. We are developing analyses to incorporate our understanding of event generation and attempts to model the selective pressures on each genetic element within regions of interest. This model should incorporate, the likelihood of a given set of

events based on the location of driver genes as well as other, more common genetic elements, such as essential genes, toxic genes, spacing of genetic elements within a particular region. We are also trying to better account for variations in levels of genomic disruption between samples and diseases.

Incorporation of whole genome sequencing data to improve models of SCNA determination

While SNP6.0 arrays are constantly being analyzed in more significant ways to produce richer data, there will always be fundamental limitations to the technology. Unlike sequencing approaches, arrays do not provide any direct information about somatic genomic structural changes, such as translocations. This necessitates a rather simplistic view of SCNAs, where each event affects contiguous regions of DNA (with minor extensions such as those described in in chapter 2). We know this model to be inaccurate. Fortunately, whole genome sequencing is being applied to more and more tumors. By comparing our event determination from SNP arrays to event decomposition techniques utilizing WGS, we hope to improve our methods of event deconstruction in non-WGD samples, and thus extend the information gained through WGS to the thousands of samples that have been genomically characterized without WGS.

Extracting functional information from correlative structure in the cancer genome

We are currently working on extending our SCNA correlation analysis to incorporate discrete, allelic information, as well as exomic sequencing data. In these analyses, we are controlling for sample specific rates of mutations in a similar way to our controls for SCNA rates in each sample. Activation or inactivation of driver genes may be accomplished in many ways during cancer development, so correlation

analyses integrating all types of alterations will improve on these results and have the potential to be very informative. By combining whole exome sequencing with a precise determination of the amount of DNA at each locus in each sample, we can generate a model of expected mutation distribution and develop a permutation test that maintains this distribution, in a way similar to our SCNA method.

In addition, we are controlling for the ploidy at each locus, as we expect that will affect the probability of a mutation at that locus. Finally, it has been observed that some specific alterations, such as *TP53* or *PIK3CA*, actually correlate or anticorrelate with overall levels of disruption above and beyond the expected rate of correlation[261]. We are attempting to control for these locus-specific patterns of alteration as well. This will lead to an ability to determine how mutations and SCNAs interact with each other in a locus and sample specific manner.

Varying levels of genomic disruption across cancers are likely to engender biases in analyses of correlations not only between SCNAs, but also between SCNAs and other features of these cancers. For example, increased genomic disruption has been associated with poor prognosis in multiple cancer types[8,363]. Poor prognosis is therefore likely to be associated with increased rates of SCNA across much of the genome. Controlling for this tendency will be required to identify SCNAs that are functionally associated with progression. It will also be important to account for other possible confounders, such as mechanistically linked events (e.g. chromothripsis or SCNAs that encompass multiple peak regions).

The future of CYCLOPS

Identification of novel non-driver targets of cancer dependency

It is important to note that the initial pooled RNAi screen is a hypothesis-generating experiment, and that, for a number of experimental reasons, there is likely a high false negative rate in our initial analysis that suggests our 56 initial candidate genes we identified may be an underestimate of the true number of potential targets. Indeed, more recent analyses using around 200 Achilles-screened cell lines (as opposed to the original 86) have identified over significant 300 genes in total, with extra genes mostly derived from an increase in power. Our initial analysis of Project Achilles only included 5,312 genes and many of these genes may represent false negative results due to insufficiently effective shRNAs. The set of 86 cell lines was also too small to enable detection of lineage-specific CYCLOPS genes. Indeed, a second RNAi dataset enriched in breast and pancreatic lineages revealed new CYCLOPS targets in addition to validating the targets described in our more lineage diverse dataset. Systematic evaluation of the complete annotated genome using more shRNAs for each gene and a larger group of cell lines representing many lineages is likely to uncover many more potential targets.

Resistance to Cyclops. Experiments with clonal expansion after transfection.

To determine the mechanisms of potential resistance to CYCLOPS vulnerabilities, selecting and characterizing single cell clones of *PSMC2*^{Loss} lines that survive *PSMC2* inhibition seems the most natural first step. To find resistances that are biologically relevant, we can rule out failures in our perturbation system, such as loss of Dox promoter or deletions that remove the effectiveness of the shRNA, by confirming Dox addition still leads to *PSMC2* inhibition. Assuming resistance is not developed by circumventing mRNA suppression, *PSMC2* qPCR, rtPCR, SDS-PAGE and native gel

western blots may provide insights into resistance mechanisms. One likely mechanism is increased expression, either through copy number amplification or increased transcription of the remaining allele, either of which could be validated by qPCR and or FISH.

Challenges of interpreting RNAi pooled screens

shRNAs are an imperfect tool for the discovery of haplo-sufficient bystander dependencies. Even in a best case scenario, in order for a gene to be correctly identified as a CYCLOPS gene in our screen, the following criteria had to have been met:

1. It is a CYCLOPS gene
2. It has two shRNA in the screen that work
3. Those two shRNA have to both have to have incomplete suppression, yet sufficient suppression to induce the phenotype in partial loss cells.
4. We have enough cell lines within and across lineages with partial loss at the locus to adequately assess the significance in gene dependency differences between loss and neutral lines.

Given these restrictions, it is perhaps surprising that our initial analysis identified as many genes as it did, and as these pooled screens add more shRNA per gene, more gene targets, more cell lines, and better analytics, we may be able to decrease the number of false negatives caused by these experimental limitations.

Regardless, the correlations we found between dependency on CYCLOPS candidates and partial loss in the context of a pooled screen are only hypotheses until they are further validated. Each would each need to be experimentally studied to

confirm our genome-scale results as interpretation of shRNA data is challenging given the preponderance of variable efficacy and off-target effects. In our initial validation, we used a targeted, low throughput approach towards validation, but more "medium-throughput" approaches, including using pooled cell culture technologies, may improve our ability to validate CYCLOPS sensitivities.

Other forms of non-driver dependency

Our studies using the Achilles cell line screen discovered a set of gene dependencies based on partial copy loss of that gene. Besides copy-number loss, other types of genomic alteration may also unveil CYCLOPS-related vulnerabilities. In most cases, vulnerability to suppression of CYCLOPS genes was associated with decreased expression, so other somatic genetic events, such as sequence variants, epigenetic modification or chromosome translocations, may similarly predict dependency. While many CYCLOPS genes exhibited high correlation between copy number and expression, many others did not, and in general, it may be perilous to assume a correlation between low expression and increased vulnerability across a panel of cell lines will extrapolate to *in vivo* dependency. For example, cell lines derived from a specific cell lineage may have slightly lower expression of a gene, yet still be dependent on it, so expression alone may be dangerous to use as a criteria for bystander vulnerability.

There is also potential for utilizing trans-effects in bystander gene loss. Cancers often rely on redundant pathways to make up for acquired deficiencies. For example, cancers with BRCA1 loss are more dependent on PARP, a second DNA repair pathway[364]. Similarly, Muller et al. showed that homozygous deletion of the glycolytic

gene enolase ENO1, a bystander gene in GBM, will leave cells more dependent on ENO2, which may prove to be a therapeutic target[365]. Although further work will be necessary to explore these other classes, CYCLOPS genes alone or in combination with oncogenic targets may provide a new approach to cancer therapy.

Pharmacologic targeting of CYCLOPS genes

One of the primary motivations for searching for vulnerabilities in bystander alterations is to increase our library of targetable dependencies. Unfortunately, many of our CYCLOPS candidates do not come from traditionally targetable gene classes. Pairing genetics with small molecule screens could highlight passenger vulnerabilities that are responsive to currently available small molecules. Unlike shRNA screens, it may be more challenging to determine whether a drug dependency predicted by bystander loss is caused by direct inhibition of the gene or through a *trans* effect. For example, lenalidomide, used to treat multiple myeloma, has recently been shown to decrease abundance of two transcription factors by increasing effectiveness of their ubiquitin-mediated degradation. One could imagine finding similar mechanisms to either target CYCLOPS genes with current drugs or discover new bystander targets through small molecule screens.

Mechanism of dependency on CYCLOPS genes

While the biochemistry of proteasome assembly has been recently elucidated, the regulation and limiting factors of 19S assembly is less well understood. We found that there is relatively little compensation for moderate suppression of one component of the proteasome, either transcriptionally or translationally. In the case of partial suppression of *PSMC2*, this leads to decreased levels of the immediate PSMC2-

containing subcomplex, but relatively stable amounts 19S and 20S proteasome. However, suppression of other genes within the complex may lead to very different biochemical results, including significant loss in proteasome function, or even inhibition of the negative feedback loop that regulates transcription of 26S proteasome components. Determining the transcriptional, translational, and biochemical responses to loss of these other genes may improve our understanding of how the proteasome is regulated and what triggers changes in complex formation.

The next step: Expanding our understanding of the CYCLOPS phenomenon

While our studies on PSMC2 showed how partial loss may lead to increased sensitivity, it is unlikely to be the only mechanism for CYCLOPS dependence. In fact, some of our CYCLOPS genes showed little correlation between copy number and expression, while others undoubtedly have other mechanisms of dosage compensation. To that end, researching the mechanism of sensitivity in other CYCLOPS genes may lead to improved understanding of this class of genes, in addition to new therapeutic targets.

The second-most enriched pathway in our CYCLOPS gene analysis was the spliceosome. Like the proteasome, the spliceosome is a cellular essential macromolecular complex[366]. The SF3B subcomplex comprises a part of the spliceosome important in splice-site recognition, and this complex alone contained multiple significance CYCLOPS genes. SF3B1 is an obligate member of this complex and one of the top CYCLOPS genes in an updated analysis using an increased number of Achilles lines. Unlike *PSMC2* or its immediate subcomplex, there are small molecule inhibitors of the SF3B complex, making partial loss of SF3B1 an enticing therapeutic

target. Validation and biochemical characterization experiments on *SF3B1*^{Loss} lines are currently underway, and initial results are promising, though the mechanism of sensitivity in loss lines may be more complicated than in the case of PSMC2. Besides the immediate potential for CYCLOPS to predict small molecule sensitivity, these experiments further validate CYCLOPS as a novel class of cancer vulnerabilities.

New high-throughput techniques may help identify inaccuracies in our current screen

Since our initial validation of PSMC2 as a CYCLOPS gene, a pathway utilized by prokaryotes as an immune response to viri, nicknamed CRISPR, has been engineered as a new tool for genetic perturbation[367]. CRISPR works by using a guide RNA strand to direct the CAS9 enzyme, a nuclease, to that sequence in the genome, where it then induces a double stranded break. Resolution of these double stranded breaks are error prone, so this system allows for directed genome modification by insertion/deletion. It is believed that these CRISPR techniques suffer fewer of the specificity/sensitivity issues that plague shRNA experiments, and so provide a much cleaner system for studying the phenotypic effects of loss-of-function[230]. Because many CYCLOPS candidates are thought to be essential genes, this is an interesting tool for creating an isogenic system with inactivation of a single allele in the CYCLOPS gene (as biallelic inactivation is unlikely to be tolerated). In fact, for studies of SF3B1, our lab has created just such an isogenic system and shown that these newly created *SF3B1*^{Loss} created lines are newly sensitive to SF3B1 suppression.

Conclusion

Somatic copy number alterations are fundamental part of cancer evolution, occur through a variety of processes, and result in a complex pattern of changes that span the

genome. Our work endeavored to improve the understanding of SCNAs and their consequence on cancer development and vulnerabilities. Through this work, we have developed new ways to analyze the patterns and significance of SCNAs throughout cancer, as well as provide a valuable resource for future studies.

The extended nature of SCNAs allowed us to query whether affected non-driver genes may introduce new opportunities for targeted therapy. Our studies of CYCLOPS genes suggest that non-driver dependencies may represent an underexploited source of cancer vulnerability. We showed that in the case of *PSMC2*, partial loss led to decreased mRNA and protein, but not a decrease in function, leaving cells with loss more vulnerable to further insult. We hope these efforts will expand the thinking about potential therapeutic targets in cancer therapy and lead to a more comprehensive search for viable treatment options.

Appendix 1: SNP6.0 array processing

Standard Affymetric SNP6.0 Array preparation and analysis

For a period of about 5 years, The Cancer Genome Atlas (TCGA), the cancer cell line encyclopedia, and many other high-throughput genomics platforms made use of SNP 6.0 Affymetrix microarrays[144]. These arrays probe the DNA content of human samples using 906,600 single nucleotide polymorphisms (SNPs) and over 946,000 probes designed to detect copy number variation.

The SNP probes are derived from previous versions of Affymetrix arrays[143] (482,000 curated from dbSNP database[368]) in addition to 424,000 SNPs discovered in the International HapMap project[369]. Each of the 2 alleles at each locus are measured in triplicate by 25-mers scattered across the array.

In addition to the SNP probes, there are 946,000 25-mer copy number (CN) probes included in SNP6.0 arrays, 744,000 were chosen to provide even coverage of the genome, with 202,000 chosen based on known germline copy number variation[370]. An overview of the bench procedure is diagrammed in [143]. The DNA sample is digested with restriction enzymes Sty 1 and NSP 1, then ligated to adapters designed to match the single stranded DNA overhangs that result from that digestion. A primer that recognizes the adapter sequence is then used to amplify the DNA through polymerase chain reaction (PCR), with settings set to optimize amplification of fragments between 200 and 1,100 bp. The resulting product is then fragmented using DNAse1 and chemically labeled with a fluorescent dye. Finally, each reaction is loaded onto a

SNP6.0 array. These arrays are scanned using a GeneChip Scanner 3000 7G and controlled by either Affymetrix GeneChip Command Console (AGCC) or GeneChip Operating Software (GCOS), either of which output a standard .CEL file that contains information about the intensity, standard deviation, and pixel count for each cell in the array. This .CEL file represents the endpoint of standard Affymetrix analyses, to which developers at the Broad Institute have developed a number of additional tools to extract information about Somatic copy number changes in cancer samples.

Broad institute Copy number inference pipeline for SNP 5.0 and 6.0 arrays[371]

Preprocessing

The first step in this pipeline is to normalize each array to correct for overall differences in array intensity, and then convert each probeset representing a SNP allele to a single value.

Median normalization adjusts the median probe-level value of each array to 1000, followed by quantile normalization. Next we use Model-based expression indices (MBEI) to map these normalized intensities to the normal sample with total intensity closest to the median total intensity in the plate. MBEI assumes a linear relationship between probe intensity and DNA content,such that the coefficient for this relationship may be unique to the probe in question. Finally, replicate probes are summarized using median polish across the samples in the plate.

Copy Number Inference

The goal of this section is to convert these normalized intensities (which are relative values based on co-developed samples) to copy number values, values relative to the total amount of DNA at a particular locus. We assume the intensity for the i-th probeset in the j-th sample, y_{ij}, is derived from a linear transformation of the underlying copy number

$$y_{ij} = \beta_{i0} + \beta_{i1} x_{ij}$$

where β_{i0} denotes the "background" parameter and β_{i1} denotes the scale parameter. Because of extra information inherent in SNP probes, the pipeline handles SNP and CN probes slightly differently

CN probes

CN probe calibration utilizes the intensity measurements of 5 cell lines, each with a different number of copies of the X chromosome (1-5). This panel of cell lines was used to create a X-chromosome "dosage" experiment, where the normalized intensity values for each probe on the X chromosome was measured for each cell line in the panel. These values were then used to fit a probe-specific linear calibration curve for each of these probes. To extrapolate these curves to the autosomal chromosomes, a regression model was used, with probe-specific GC content, fragment length, and median intensity used as regression variables over this set of 5 cell lines (presumably diploid over the autosomal chromosomes). Finally, this regression model is used to

predict the background and scale parameters on each probe in the current dataset (**Table 1**).

SNP probes

The background and scale parameters for SNP probes rely on the Birdseed, which calculates these parameters by analyzing relative intensities of zero and one copy SNP varients in diploid “normals” (non-tumor somatic tissue, usually derived from patient blood) run in the same batch. Because this relies on representation of the allele within the batch of normal, it is optimal to run at least 15 diploid normal samples in each batch. However, if a batch lacks adequate representation at a particular locus we can use prior information to fill in these values. Otherwise, the probes are discarded for this set of samples.

Sample cleanup and noise reduction

Outlier removal

Noise is a constant factor in analysis of DNA arrays. This noise can be sample specific (e.g. a bubble on the array), or systematic (e.g. inefficient DNA cleavage at a particular locus). Outlier elimination attempts to remove probes with extreme values not corroborated by other nearby probes. Specifically, taking the 5 probes immediately to one side of a given probe, if the difference between the given probe and the median of the other 5 probes is greater than 0.3, the probe is considered an outlier with respect to

these probes. If a probe is an outlier with respect to probes on both its left, and its right, it is replaced with the median of the three values centered on itself.

Tangent normalization

Even after these extensive normalization procedures, there is variation in probe set intensity across samples that remain when comparing individual diploid normal samples, or even replicates of the same normal sample. These patterns of variation may reflect changes in experimental conditions between different arrays but could provide false copy number alterations in not properly accounted for. To remove these variations caused by systemic noise present in normal tissue, we project the copy number values of each probe in our tumor sample onto a plane created by values obtained from a large archive of normal tissue (presumably diploid). The linear projection of our sample onto this "plane" of normal DNA values represents the hypothetical diploid sample subjected to a similar set of systemic noise as our tumor sample. We then subtract this modeled systemic variation from our tumor sample. It is important to point out that known regions of germline copy number variation (CNVs) are removed prior to this step as this is an attempt to model systematic (i.e. non-biological) noise alone. CNV regions are subsequently added back to data.

Circular Binary Segmentation (CBS) [141,372]

This approach takes each chromosome and attempts to find evidence for contiguous regions of equal copy number and transforms the data from thousands of

probes, each with a unique copy number value, to segments of data, where each segment is composed of multiple probes with a single copy number value.

CBS does this by taking each segment (starting with the whole chromosome) and recursively asking if there is likely to be a copy number breakpoint within that segment.

If a segment consists of markers 1 through m, this equates to asking at each marker pair *i* and *j (i<j),* whether the mean of values *i+1*:j is significantly different from the mean of values outside of this range (**Figure 2**). This significance is determined by a "hybrid" approach. If the number of markers *m* in a given segment is less than 1000, then the maximum value k for which full permutation test is applied is 25, with k increasing by 5 for every doubling of *m.* For segments smaller than this threshold, the p-value is derived directly from a full permutation test for all values of *i* and *j* such that *(j – i)* > than some minimum segment length (based on inter-probe variation within the sample). For values of (*j – i)* greater than the threshold, the test statistic *T* was assumed to originate from a standard normal distribution. If there exists an (*i,j)* pair that exceeds the significance threshold, the (*i,j)* that generate the best *T* statistic are considered the end points of a new segment. This process is repeated recursively on each segment, until no more segments are created by this approach. Finally, the mean value of markers in each segment becomes the value of that segment.

HAPSEG and ABSOLUTE

The absolute algorithm is a parallel analysis technique developed by Carter et al[147] utilizing the extra information present in SNP probes to determine integer allelic copy number values at each locus.

References

1. Boveri, T. Concerning the Origin of Malignant Tumours by Theodor Boveri. Translated and annotated by Henry Harris. *Journal of Cell Science* **121,** 1–84 (2008).
2. Hansemann, von, D. Ueber asymmetrische Zelitheilung in epithel Krebsen und deren biologische Bedeutung. *Virchow's Arch. Path. Anat.* (1890).
3. Nowell, P.C. & Hungerford, D. A. A minute chromosome in human chronic granulocytic leukemia. *Science* (1960).
4. Knudson, A. G., Meadows, A. T., Nichols, W. W. & Hill, R. Chromosomal Deletion and Retinoblastoma. *N. Engl. J. Med.* **295,** 1120–1123 (1976).
5. Baker, S., Fearon, E., Nigro, J. & Vogelstein, B. Chromosome 17 Deletions and p53 gene Mutations in colorectal carcinomas. *Science* **244,** 217 (1989).
6. Cavenee, W. K., Dryja, T. P., Phillips, R. A. & White, R. L. Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. *Nature* **305,** 779–784 (1983).
7. Hayward, W. S., Neel, B. G. & Astrin, S. M. Activation of a cellular onc gene by promoter insertion in ALV-induced lymphoid leukosis. *Nature* **290,** 475–480 (1981).
8. Pinto, A. E. *et al.* DNA Ploidy is an Independent Predictor of Survival in Breast Invasive Ductal Carcinoma: A Long-term Multivariate Analysis of 393 Patients. *Annals of Surgical Oncology* **20,** 1530–1537 (2013).
9. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* **339,** 1546–1558 (2013).
10. Futreal, P. A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4,** 177–183 (2004).
11. Nowell, P.C. *Science* **194,** 23–28 (1976).
12. Merlo, L. M. F., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nat Rev Cancer* **6,** 924–935 (2006).
13. GOODMAN, L. S., WINTROBE, M. M., Dameshek, W., Goodman, M. J. & Gilman, A. Nitrogen Mustard Therapy - Use of Methyl-Bis(Beta-Chloroethyl)Amine Hydrochloride and Tris(Beta-Chloroethyl)Amine Hydrochloride for Hodgkins Disease, Lymphosarcoma, Leukemia and Certain Allied and Miscellaneous Disorders. *Journal of the american medical association* **132,** 126–132 (1946).
14. Sudhakar, A. History of Cancer, Ancient and Modern Treatment Methods. *J Canc Sci Ther* **01,** i–iv (2009).
15. KIRSCHSTEIN, R. L. & GERBER, P. Ependymomas produced after intracerebral

inoculation of SV40 into new-born hamsters. *Nature* **195,** 299–300 (1962).
16. Martin, G. S. The hunting of the Src. *Nat Rev Mol Cell Biol* **2,** 467–475 (2001).
17. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100,** 57–70 (2000).
18. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674 (2011).
19. Lawrence MS, S. P. P. P. K. G. C. K. S. A. C. S. S. C. M. C. R. S. K. A. H. P. M. A. D. Y. Z. L. R. A. P. T. S. N. H. E. K. J. S. C. A. L. N. E. S. E. C. M. A. D. S. G. V. D. N. M. D. D. L. P. L. L. H. D. F. T. I. M. H. B. H. E. B. S. D. A. L. J. L. D. W. C. M.-Z. J. H.-M. A. K. A. M. S. M. J. L. R. C. B. O. R. P. M. W. W. A. K. G. S. R. C. B. J. S. K. B. A. G. L. M. M. G. T. G. D. S. S. L. E. G. G. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **489,** (2013).
20. Kerr, J. F., Wyllie, A. H. & Currie, A. R. Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *British journal of cancer* **26,** 239–257 (1972).
21. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45,** 1134–1140 (2013).
22. Murphree, A. L. & Benedict, W. F. Retinoblastoma: clues to human oncogenesis. *Science* **223,** 1028–1033 (1984).
23. *Where do the millions of cancer research dollars go every year*. (2013). at <http://www.slate.com/blogs/quora/2013/02/07/where_do_the_millions_of_cancer_research_dollars_go_every_year.html>
24. Eckhouse, S. & Sullivan, R. A Survey of Public Funding of Cancer Research in the European Union. 1–6 (2006). doi:10.1371/journal
25. *Advances Elusive in the Drive to Cure Cancer*. (2009). at <http://www.nytimes.com/2009/04/24/health/policy/24cancer.html?_r=0>
26. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2011).
27. Bignell, G. R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463,** 893–898 (2011).
28. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).
29. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat Rev Genet* **10,** 551–564 (2009).
30. Zhang, C. Z., Leibowitz, M. L. & Pellman, D. Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. *Genes & development* **27,** 2513–2530 (2013).
31. Bunting, S. F. & Nussenzweig, A. End-joining, translocations and cancer. 1–12 (2013). doi:10.1038/nrc3537
32. Inaki, K. & Liu, E. T. Structural mutations in cancer: mechanistic and functional insights. *Trends in Genetics* **28,** 550–559 (2012).
33. Stephens, P. J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462,** 1005–1010 (2009).
34. Weir, B. A. *et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450,** 893–898 (2007).
35. Network, T. C. G. A. R. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–525 (2012).

36. Xue, W. *et al.* A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions. **109,** 8212–8217 (2012).
37. Nik-Zainal, S. *et al.* The Life History of 21 Breast Cancers. *Cell* **149,** 994–1007 (2012).
38. Tsao, M. S. *et al.* Erlotinib in lung cancer - Molecular and clinical predictors of outcome. *N. Engl. J. Med.* **353,** 133–144 (2005).
39. Howell, S. J., Wardley, A. M. & Armstrong, A. C. Re: Ki67 Index, HER2 Status, and Prognosis of Patients With Luminal B Breast Cancer. *JNCI Journal of the National Cancer Institute* **101,** 1730–1730 (2009).
40. Kim, E. S. *et al.* Gefitinib versus docetaxel in previously treated non-small-celllung cancer (INTEREST): a randomised phase III trial. *Lancet* **372,** 1809–1818 (2008).
41. Lowe, S. W. *et al.* p53 status and the efficacy of cancer therapy in vivo. *Science* **266,** 807–810 (1994).
42. DeVita, V. T., Jr. & Rosenberg, S. A. Two Hundred Years of Cancer Research. *N. Engl. J. Med.* **366,** 2207–2214 (2012).
43. Hajdu, S. I. A note from history: Landmarks in history of cancer, part 1. *Cancer* **117,** 1097–1102 (2010).
44. Marie, P., Clunet, J. & Raviot-LaPointe, G. Contribution a L'etude du development des tumeurs malignes sur le ulcers de Roentgen. *Bull Assoc Franc l'etude de Cancer* 404–426 (1910).
45. Hajdu, S. I. & Darvishian, F. A note from history: Landmarks in history of cancer, part 5. *Cancer* **119,** 1450–1466 (2013).
46. Hajdu, S. I. & Vadmal, M. A note from history: Landmarks in history of cancer, Part 6. *Cancer* **119,** 4058–4082 (2013).
47. Muller, H. J. ARTIFICIAL TRANSMUTATION OF THE GENE. *Science* **66,** 84–87 (1927).
48. Auerbach, C., Robson, J. M. & Carr, J. G. The Chemical Production of Mutations. *Science* **105,** 243–247 (1947).
49. Linder, D. & Gartler, S. M. Glucose-6-phosphate dehydrogenase mosaicism: utilization as a cell marker in the study of leiomyomas. *Science* **150,** 67–69 (1965).
50. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 20007–20012 (2007).
51. Taylor, B. S. *et al.* Functional copy-number alterations in cancer. *PLoS ONE* **3,** e3179 (2008).
52. Krasnitz, A., Sun, G., Andrews, P. & Wigler, M. Target inference from collections of genomic intervals. *Proc. Natl. Acad. Sci. U.S.A.* **110,** E2271–8 (2013).
53. Ford, C. E. & Hamerton, J. L. A COLCHICINE, HYPOTONIC CITRATE, SQUASH SEQUENCE FOR MAMMALIAN CHROMOSOMES. *Stain Technology* **31,** 247–251 (1956).
54. Ford, C. E. & Hamerton, J. L. The chromosomes of man. *Nature* **178,** 1020–1023 (1956).
55. Atkin, N. B. & RICHARDS, B. M. Deoxyribonucleic acid in human tumours as measured by microspectrophotometry of Feulgen stain: a comparison of

tumours arising at different sites. *British journal of cancer* **10,** 769–786 (1956).
56. Nowell, P.C. & Hungerford, D. A. Chromosome studies on normal and leukemic human leukocytes. *J. Natl. Cancer Inst.* **25,** 85–109 (1960).
57. Atkin, N. B., Mattinson, G. & Baker, M. C. A comparison of the DNA content and chromosome number of fifty human tumours. *British journal of cancer* **20,** 87–101 (1966).
58. WAKONIG-VAARTAJA, R. Mutation as initiator of neoplasms, and their secondary evolution. *Nature* **193,** 144–145 (1962).
59. Nowell, P.C. & Hungerford, D. A. CHROMOSOME CHANGES IN HUMAN LEUKEMIA AND A TENTATIVE ASSESSMENT OF THEIR SIGNIFICANCE. *Ann. N. Y. Acad. Sci.* **113,** 654–662 (1964).
60. Caspersson, T. *et al.* Chemical differentiation along metaphase chromosomes. *Exp. Cell Res.* **49,** 219–222 (1968).
61. Caspersson, T. *et al.* DNA-binding fluorochromes for the study of the organization of the metaphase nucleus. *Exp. Cell Res.* **58,** 141–152 (1969).
62. Caspersson, T., Zech, L., Johansson, C. & Modest, E. J. Identification of human chromosomes by DNA-binding fluorescent agents. *Chromosoma* **30,** 215–227 (1970).
63. Arrighi, F. E. & Hsu, T. C. Localization of heterochromatin in Human Chromosomes. *Cytogenetics* **10,** 81–& (1971).
64. Hsu, T. C. & Arrighi, F. E. Distribution of constitutive heterochromatin in mamallian chromosomes. *Chromosoma* **34,** 243–253 (1971).
65. Drets, M. E. & Shaw, M. W. Specific banding patterns of human chromosomes. *Proc. Natl. Acad. Sci. U.S.A.* **68,** 2073–2077 (1971).
66. Finaz, C. & de Grouchy, J. Identification of individual chromosomes in the human karyotype by their banding pattern after proteolytic digestion. *Humangenetik* **15,** 249–252 (1972).
67. Hamerton, J. L., Jacobs, P. A., Klinger, H. P. & Geffner, E. S. Paris Conference (1971): Standardization in human cytogenetics. *Cytogenetics* **11,** 317–362 (1972).
68. Knudson, A. G., Meadows, A. T., Nichols, W. W. & Hill, R. Chromosomal Deletion and Retinoblastoma. *JAMA ophthalmology* **82,** 1–4 (1969).
69. Rowley, J. D. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243,** 290–293 (1973).
70. Francke, U., Holmes, L. B. & Riccardi, V. M. Aniridia-Wilms' tumor association: evidence for specific deletion of 11p13
. *Cytogenetics and cell genetics* **24,** 185–192 (1979).
71. Rowley, J. D. & Testa, J. R. Chromosome Abnormalities in Malignant Hematologic Diseases
. *Advances in cancer research* **36,** 103–148 (1982).
72. Rowley, J. D. Human oncogene locations and chromosome aberrations. *Nature* **301,** 290–291 (1983).
73. Mitelman, F. & Levan, G. Clustering of aberrations to specific chromosomes in human neoplasms. IV. A survey of 1,871 cases. *Hereditas* **95,** 79–139 (1981).
74. Manolov, G. & Manolova, Y. Marker band in one chromosome 14 from Burkitt

lymphomas. *Nature* **237,** 33–34 (1972).
75. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **74,** 5463–5467 (1977).
76. Hewick, R. M., Hunkapiller, M. W., Hood, L. E. & Dreyer, W. J. A gas-liquid solid phase peptide and protein sequenator. *Journal of Biological Chemistry* **256,** 7990–7997 (1981).
77. Driscoll, D. A. *et al.* Deletions and microdeletions of 22q11.2 in velo-cardio-facial syndrome. *Am. J. Med. Genet.* **44,** 261–268 (1992).
78. Chieffo, C. *et al.* Isolation and characterization of a gene from the DiGeorge chromosomal region homologous to the mouse Tbx1 gene. *Genomics* **43,** 267–277 (1997).
79. Hall, J. M. *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250,** 1684–1689 (1990).
80. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266,** 66–71 (1994).
81. Oren, M. & Levine, A. J. Molecular cloning of a cDNA specific for the murine p53 cellular tumor antigen. *Proc. Natl. Acad. Sci. U.S.A.* **80,** 56–59 (1983).
82. Zakut-Houri, R. *et al.* A single gene and a pseudogene for the cellular tumour antigen p53. *Nature* **306,** 594–597 (1983).
83. Brugge, J. S. & Erikson, R. L. Identification of a transformation-specific antigen induced by an avian sarcoma virus. *Nature* **269,** 346–348 (1977).
84. Shih, C., Shilo, B. Z., Goldfarb, M. P., Dannenberg, A. & Weinberg, R. A. Passage of phenotypes of chemically transformed cells via transfection of DNA and chromatin. *Proc. Natl. Acad. Sci. U.S.A.* **76,** 5714–5718 (1979).
85. Der, C. J., Krontiris, T. G. & Cooper, G. M. Transforming genes of human bladder and lung carcinoma cell lines are homologous to the ras genes of Harvey and Kirsten sarcoma viruses. *Proc. Natl. Acad. Sci. U.S.A.* **79,** 3637–3640 (1982).
86. Cooper, G. M. & Neiman, P. E. Transforming genes of neoplasms induced by avian lymphoid leukosis viruses. *Nature* **287,** 656–659 (1980).
87. Lane, D. P. & Crawford, L. V. T antigen is bound to a host protein in SV40-transformed cells. *Nature* **278,** 261–263 (1979).
88. Matlashewski, G. *et al.* Isolation and characterization of a human p53 cDNA clone: expression of the human p53 gene. *Embo J.* **3,** 3257–3262 (1984).
89. Erikson, J., ar-Rushdi, A., Drwinga, H. L., Nowell, P.C. & Croce, C. M. Transcriptional activation of the translocated c-myc oncogene in burkitt lymphoma. *Proc. Natl. Acad. Sci. U.S.A.* **80,** 820–824 (1983).
90. Downward, J. *et al.* Close similarity of epidermal growth factor receptor and v-erb-B oncogene protein sequences. *Nature* **307,** 521–527 (1984).
91. Galibert, F., Mandart, E., Fitoussi, F., Tiollais, P. & Charnay, P. Nucleotide sequence of the hepatitis B virus genome (subtype ayw) cloned in E. coli. *Nature* **281,** 646–650 (1979).
92. Orth, G. *et al.* Characterization of two types of human papillomaviruses in lesions of epidermodysplasia verruciformis. *Proc. Natl. Acad. Sci. U.S.A.* **75,** 1537–1541 (1978).
93. Correa, P. *et al.* Helicobacter pylori and gastric carcinoma. Serum antibody prevalence in populations with contrasting cancer risks. *Cancer* **66,** 2569–2574

(1990).
94. Collins, F. S. Positional cloning: Let's not call it reverse anymore. *Nat Genet* **1,** 1–4 (1992).
95. Orkin, S. H. Reverse genetics and human disease. *Cell* **47,** 845–850 (1986).
96. Francke, U. *et al.* Minor Xp21 chromosome deletion in a male associated with expression of Duchenne muscular dystrophy, chronic granulomatous disease, retinitis pigmentosa, and McLeod syndrome. *Am. J. Hum. Genet.* **37,** 250–267 (1985).
97. van Ommen, G. J. *et al.* A physical map of 4 million bp around the Duchenne muscular dystrophy gene on the human X-chromosome. *Cell* **47,** 499–504 (1986).
98. van Heyningen, V. *et al.* Molecular analysis of chromosome 11 deletions in aniridia-Wilms tumor syndrome. *Proc. Natl. Acad. Sci. U.S.A.* **82,** 8592–8596 (1985).
99. Friend, S. H. *et al.* A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323,** 643–646 (1986).
100. Burke, D. T., Carle, G. F. & Olson, M. V. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236,** 806–812 (1987).
101. Lovett, M., Kere, J. & Hinton, L. M. Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proc. Natl. Acad. Sci. U.S.A.* **88,** 9628–9632 (1991).
102. Parimoo, S., Patanjali, S. R., Shukla, H., Chaplin, D. D. & Weissman, S. M. cDNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* **88,** 9623–9627 (1991).
103. DeLisi, C. Santa Fe 1986: Human genome baby-steps. *Nature* **455,** 876–877 (2008).
104. Pollack, J. R. *et al.* Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* **23,** 41–46 (1999).
105. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470 (1995).
106. Bowtell, D. D. Options available--from start to finish--for obtaining expression data by microarray. *Nat Genet* **21,** 25–32 (1999).
107. Shalon, D., Smith, S. J. & Brown, P. O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* **6,** 639–645 (1996).
108. DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* **14,** 457–460 (1996).
109. Lashkari, D. A. *et al.* Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. U.S.A.* **94,** 13057–13062 (1997).
110. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U.S.A.* **96,** 2907–2912 (1999).

111. Golub, T. R. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* **286,** 531–537 (1999).
112. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* **20,** 207–211 (1998).
113. Fodor, S. P. *et al.* Light-directed, spatially addressable parallel chemical synthesis. *Science* **251,** 767–773 (1991).
114. Fodor, S. P. *et al.* Multiplexed biochemical assays with biological chips. *Nature* **364,** 555–556 (1993).
115. Pease, A. C. *et al.* Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. U.S.A.* **91,** 5022–5026 (1994).
116. Chee, M. *et al.* Accessing genetic information with high-density DNA arrays. *Science* **274,** 610–614 (1996).
117. Drmanac, R., Labat, I., Brukner, I. & Crkvenjakov, R. Sequencing of megabase plus DNA by hybridization: theory of the method. *Genomics* **4,** 114–128 (1989).
118. Drmanac, R. *et al.* DNA sequence determination by hybridization: a strategy for efficient large-scale sequencing. *Science* **260,** 1649–1652 (1993).
119. Cronin, M. T. *et al.* Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. *Hum. Mutat.* **7,** 244–255 (1996).
120. Hacia, J. G., Brody, L. C., Chee, M. S., Fodor, S. P. & Collins, F. S. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nat Genet* **14,** 441–447 (1996).
121. Lockhart, D. J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14,** 1675–1680 (1996).
122. Kallioniemi, A. *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258,** 818–821 (1992).
123. Manoir, du, S. *et al.* Detection of complete and partial chromosome gains and losses by comparative genomic in situ hybridization. *Hum. Genet.* **90,** 590–610 (1993).
124. Bentz, M., Plesch, A., Stilgenbauer, S., Dohner, H. & Lichter, P. Minimal sizes of deletions detected by comparative genomic hybridization. *Genes, chromosomes & cancer* **21,** 172–175 (1998).
125. Tanner, M. M. *et al.* Increased copy number at 20q13 in breast cancer: defining the critical region and exclusion of candidate genes. *Cancer Research* **54,** 4257–4260 (1994).
126. Bärlund, M. *et al.* Increased copy number at 17q22-q24 by CGH in breast cancer is due to high-level amplification of two separate regions. *Genes, chromosomes & cancer* **20,** 372–376 (1997).
127. Solinas-Toldo, S. *et al.* Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes, chromosomes & cancer* **20,** 399–407 (1997).
128. Venter, J. C. The Sequence of the Human Genome. *Science* **291,** 1304–1351 (2001).
129. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
130. Barrett, M. T. *et al.* Comparative genomic hybridization using oligonucleotide

microarrays and total genomic DNA. *Proc. Natl. Acad. Sci. U.S.A.* **101,** 17765–17770 (2004).
131. Hansen, M. F. & Cavenee, W. K. Genetics of cancer predisposition. *Cancer Research* **47,** 5518–5527 (1987).
132. Brown, M. A. Tumor suppressor genes and human cancer. *Adv. Genet.* **36,** 45–135 (1997).
133. Mei, R. Genome-wide Detection of Allelic Imbalance Using Human SNPs and High-density DNA Arrays. *Genome Res.* **10,** 1126–1137 (2000).
134. Lindblad-Toh, K. *et al.* Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nat. Biotechnol.* **18,** 1001–1005 (2000).
135. Hoque, M. O., Lee, C.-C. R., Cairns, P., Schoenberg, M. & Sidransky, D. Genome-wide genetic characterization of bladder cancer: a comparison of high-density single-nucleotide polymorphism arrays and PCR-based microsatellite analysis. *Cancer Research* **63,** 2216–2222 (2003).
136. Kennedy, G. C. *et al.* Large-scale genotyping of complex DNA. *Nat. Biotechnol.* **21,** 1233–1237 (2003).
137. Zhao, X. *et al.* Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis. *Cancer Research* **65,** 5561–5570 (2005).
138. Bignell, G. R. High-Resolution Analysis of DNA Copy Number Using Oligonucleotide Microarrays. *Genome Res.* **14,** 287–295 (2004).
139. Zhao, X. *et al.* An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Research* **64,** 3060–3071 (2004).
140. Carter, S. L., Meyerson, M. & Getz, G. Accurate estimation of homologue-specific DNA concentration ratios in cancer samples allows long-range haplotyping. *Preprint at http://precedings.nature.com/documents/6494/version/1/* (2011).
141. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. **5,** 557–572 (2004).
142. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23,** 657–663 (2007).
143. Affymetrix, Inc. 500k_datasheet. 1–4 (2006).
144. Affymetrix, Inc. genomewide_snp6_datasheet. 1–4 (2009).
145. Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40,** 1253–1260 (2008).
146. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *PNAS* **107,** 16910–16915 (2010).
147. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30,** 413–421 (2012).
148. Laird, P. W. Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet* **11,** 191–203 (2010).
149. Rabbitts, T. H. Commonality but Diversity in Cancer Gene Fusions. *Cell* **137,**

391–395 (2009).
150. Nigg, E. Centrosome aberrations: Cause or consequence of cancer progression. *Nat Rev Cancer* **3,** 815–815 (2002).
151. Bakhoum, S. F., Thompson, S. L., Manning, A. L. & Compton, D. A. Genome stability is ensured by temporal control of kinetochore–microtubule dynamics. *Nature cell biology* **11,** 27–35 (2008).
152. Meraldi, P., Lukas, J., Fry, A. M., Bartek, J. & Nigg, E. A. Centrosome duplication in mammalian somatic cells requires E2F and Cdk2-cyclin A. *Nature cell biology* **1,** 88–93 (1999).
153. Duensing, S. & Münger, K. Human papillomaviruses and centrosome duplication errors: modeling the origins of genomic instability. *Oncogene* **21,** 6241–6248 (2002).
154. Iovino, F., Lentini, L., Amato, A. & Di Leonardo, A. RB acute loss induces centrosome amplification and aneuploidy in murine primary fibroblasts. *Mol Cancer* **5,** 38 (2006).
155. Burkhart, D. L. & Sage, J. Cellular mechanisms of tumour suppression by the retinoblastoma gene. *Nat Rev Cancer* **8,** 671–682 (2008).
156. Meraldi, P., Honda, R. & Nigg, E. A. Aurora-A overexpression reveals tetraploidization as a major route to centrosome amplification in p53-/- cells. *Embo J.* **21,** 483–492 (2002).
157. Cimini, D., Wan, X., Hirel, C. B. & Salmon, E. D. Aurora Kinase Promotes Turnover of Kinetochore Microtubules to Reduce Chromosome Segregation Errors. *Current Biology* **16,** 1711–1718 (2006).
158. Cimini, D. Detection and correction of merotelic kinetochore orientation by Aurora B and its partners. *Cell Cycle* **6,** 1558–1564 (2007).
159. Rieder, C. L., Cole, R. W., Khodjakov, A. & Sluder, G. The checkpoint delaying anaphase in response to chromosome monoorientation is mediated by an inhibitory signal produced by unattached kinetochores. *J. Cell Biol.* **130,** 941–948 (1995).
160. Sironi, L. *et al.* Crystal structure of the tetrameric Mad1-Mad2 core complex: implications of a 'safety belt' binding mechanism for the spindle checkpoint. *Embo J.* **21,** 2496–2506 (2002).
161. Dobles, M., Liberal, V., Scott, M. L., Benezra, R. & Sorger, P. K. Chromosome missegregation and apoptosis in mice lacking the mitotic checkpoint protein Mad2. *Cell* **101,** 635–645 (2000).
162. Thoma, C., Toso, A., Meraldi, P. & Krek, W. Mechanisms of aneuploidy and its suppression by tumour suppressor proteins. *Swiss Med Wkly* (2011). doi:10.4414/smw.2011.13170
163. Iwanaga, Y. *et al.* Heterozygous Deletion of Mitotic Arrest-Deficient Protein 1 (MAD1) Increases the Incidence of Tumors in Mice. *Cancer Research* **67,** 160–166 (2007).
164. Cahill, D. P. *et al.* Mutations of mitotic checkpoint genes in human cancers. *Nature* **392,** 300–303 (1998).
165. Draviam, V. M., Shapiro, I., Aldridge, B. & Sorger, P. K. Misorientation and reduced stretching of aligned sister kinetochores promote chromosome missegregation in EB1- or APC-depleted cells. *Embo J.* **25,** 2814–2827 (2006).

166. Guardavaccaro, D. *et al.* Control of chromosome stability by the β-TrCP–REST–Mad2 axis. *Nature* **452,** 365–369 (2008).
167. Thoma, C. R. *et al.* VHL loss causes spindle misorientation and chromosome instability. *Nature Publishing Group* **11,** 994–1001 (2009).
168. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** (2012).
169. Torres, J. Z., Ban, K. H. & Jackson, P. K. A Specific Form of Phospho Protein Phosphatase 2 Regulates Anaphase-promoting Complex/Cyclosome Association with Spindle Poles. *Mol. Biol. Cell* **21,** 897–904 (2010).
170. Boveri, T. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. *Journal of Cell Science* **121 Suppl 1,** 1–84 (2008).
171. Hicks, J. *et al.* Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res.* **16,** 1465–1479 (2006).
172. McClintock, B. The Stability of Broken Ends of Chromosomes in Zea Mays. *Genetics* **26,** 234–282 (1941).
173. Murnane, J. P. Telomere dysfunction and chromosome instability. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **730,** 28–36 (2012).
174. Korbel, J. O. & Campbell, P. J. Criteria for Inference of Chromothripsis in Cancer Genomes. *Cell* **152,** 1226–1236 (2013).
175. Stephens, P. J. *et al.* Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell* **144,** 27–40 (2011).
176. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat Rev Genet* **10,** 551–564 (2009).
177. Liskay, R. M., Letsou, A. & Stachelek, J. L. Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells. *Genetics* **115,** 161–167 (1987).
178. Smith, C. E., Llorente, B. & Symington, L. S. Template switching during break-induced replication. *Nature* **447,** 102–105 (2007).
179. Schmidt, K. H., Wu, J. & Kolodner, R. D. Control of translocations between highly diverged genes by Sgs1, the Saccharomyces cerevisiae homolog of the Bloom's syndrome protein. *Molecular and Cellular Biology* **26,** 5406–5420 (2006).
180. Lieber, M. R. The Mechanism of Human Nonhomologous DNA End Joining. *The Journal of biological chemistry* **283,** 1–5 (2007).
181. McVey, M. & Lee, S. E. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends in Genetics* **24,** 529–538 (2008).
182. Misteli, T. & Soutoglou, E. The emerging role of nuclear architecture in DNA repair and genome maintenance. *Nat Rev Mol Cell Biol* **10,** 243–254 (2009).
183. Lukásová, E. *et al.* Localisation and distance between ABL and BCR genes in interphase nuclei of bone marrow cells of control donors and patients with chronic myeloid leukaemia. *Hum. Genet.* **100,** 525–535 (1997).
184. Wijchers, P. J. & de Laat, W. Genome organization influences partner selection for chromosomal rearrangements. *Trends Genet.* **27,** 63–71 (2011).
185. De, S. & Michor, F. DnA replication timing and long-range DnA interactions

predict mutational landscapes of cancer genomes. *Nat. Biotechnol.* **29,** 1103–1108 (2011).
186. Fudenberg, G., Getz, G., Meyerson, M. & Mirny, L. A. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat. Biotechnol.* **29,** 1109–U75 (2011).
187. de Thé, H., Chomienne, C., Lanotte, M., Degos, L. & Dejean, A. The t(15;17) translocation of acute promyelocytic leukaemia fuses the retinoic acid receptor alpha gene to a novel transcribed locus. *Nature* **347,** 558–561 (1990).
188. Aurias, A., Rimbaut, C., Buffe, D., Zucker, J. M. & Mazabraud, A. Translocation involving chromosome 22 in Ewing's sarcoma. A cytogenetic study of four fresh tumors. *Cancer Genet. Cytogenet.* **12,** 21–25 (1984).
189. Tonon, G. *et al.* High-resolution genomic profiles of human lung cancer. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 9625–9630 (2005).
190. Taylor, B. S. *et al.* Functional Copy-Number Alterations in Cancer. *PLoS ONE* **3,** (2008).
191. Bignell, G. R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463,** 893–898 (2010).
192. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).
193. Rueda, O. M. & Diaz-Uriarte, R. Finding Recurrent Copy Number Alteration Regions: A Review ofMethods. *Current Bioinformatics* **5,** (2010).
194. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology* **12,** (2011).
195. Styles, J. A. Mammalian cell transformation in vitro. Six tests for carcinogenicity. *British journal of cancer* **37,** 931–936 (1978).
196. BOYDEN, S. The chemotactic effect of mixtures of antibody and antigen on polymorphonuclear leucocytes. *J. Exp. Med.* **115,** 453–466 (1962).
197. Evan, G. A Matter of Life and Cell Death. *Science* **281,** 1317–1322 (1998).
198. Metz, T., Harris, A. W. & Adams, J. M. Absence of p53 allows direct immortalization of hematopoietic cells by the myc and raf oncogenes. *Cell* **82,** 29–36 (1995).
199. Tassone, P. *et al.* BRCA1 expression modulates chemosensitivity of BRCA1-defective HCC1937 human breast cancer cells. *British journal of cancer* **88,** 1285–1291 (2003).
200. Vandin, F., Upfal, E. & Raphael, B. J. Algorithms for Detecting Significantly Mutated Pathways in Cancer. *Journal of Computational Biology* **18,** 507–522 (2011).
201. Vandin, F., Upfal, E. & Raphael, B. J. De novo discovery of mutated driver pathways in cancer. *Genome Res.* **22,** 375–385 (2012).
202. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22,** 398–406 (2012).
203. Etemadmoghadam, D. *et al.* Synthetic lethality between CCNE1 amplification and loss of BRCA1. *PNAS* **110,** 19489–19494 (2013).
204. A, H. F. *The Pure Theory of Capital*. (1941).
205. Mitelman, F., Mertens, F. & Johansson, B. A breakpoint map of recurrent

chromosomal rearrangements in human neoplasia. *Nat Genet* **15 Spec No,** 417–474 (1997).
206. Pennisi, E. A catalog of cancer genes at the click of a mouse. *Science* **276,** 1023–1024 (1997).
207. Strausberg, R. L., Buetow, K. H., Emmert-Buck, M. R. & Klausner, R. D. The cancer genome anatomy project: building an annotated gene index. *Trends Genet.* **16,** 103–106 (2000).
208. Zhang, L. *et al.* Gene expression profiles in normal and cancer cells. *Science* **276,** 1268–1272 (1997).
209. Cheung, V. G. *et al.* Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409,** 953–958 (2001).
210. Chin, L. *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).
211. Hammerman, P. S. *et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–525 (2012).
212. Weir, B. A. *et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450,** 893–U22 (2007).
213. Brose, M. S. *et al.* BRAF and RAS mutations in human lung cancer and melanoma. *Cancer Research* **62,** 6997–7000 (2002).
214. Soda, M. *et al.* Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. *Nature* **448,** 561–566 (2007).
215. Daley, G. Q., Vanetten, R. A. & Baltimore, D. INDUCTION OF CHRONIC MYELOGENOUS LEUKEMIA IN MICE BY THE P210BCR/ABL GENE OF THE PHILADELPHIA-CHROMOSOME. *Science* **247,** 824–830 (1990).
216. Kwak, E. L. *et al.* Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N. Engl. J. Med.* **363,** 1693–1703 (2010).
217. Druker, B. J. *et al.* Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *N. Engl. J. Med.* **355,** 2408–2417 (2006).
218. Chapman, P. B. *et al.* Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. *N. Engl. J. Med.* **364,** 2507–2516 (2011).
219. Horn, L. & Pao, W. EML4-ALK: Honing In on a New Target in Non-Small-Cell Lung Cancer. *J. Clin. Oncol.* **27,** 4232–4235 (2009).
220. Koehler, A. N. A complex task? Direct modulation of transcription factors with small molecules. *Current Opinion in Chemical Biology* **14,** 331–340 (2010).
221. Hong, B., van den Heuvel, A. P. J., Prabhu, V. V., Zhang, S. & El-Deiry, W. S. Targeting tumor suppressor p53 for cancer therapy: strategies, challenges and opportunities. *Curr Drug Targets* **15,** 80–89 (2014).
222. Hahn, W. C. *et al.* Creation of human tumour cells with defined genetic elements. *Nature* **400,** 464–468 (1999).
223. Boehm, J. S., Hession, M. T., Bulmer, S. E. & Hahn, W. C. Transformation of Human and Murine Fibroblasts without Viral Oncoproteins. *Molecular and Cellular Biology* **25,** 6464–6474 (2005).
224. Hahn, W. C. & Weinberg, R. A. Rules for making human tumor cells. *N. Engl. J. Med.* **347,** 1593–1603 (2002).
225. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100,** 57–70 (2000).
226. Kaelin, W. G. The concept of synthetic lethality in the context of anticancer

therapy. *Nat Rev Cancer* **5,** 689–698 (2005).
227. Ashworth, A., Lord, C. J. & Reis-Filho, J. S. Perspective. *Cell* **145,** 30–38 (2011).
228. Solimini, N. L., Luo, J. & Elledge, S. J. Non-oncogene addiction and the stress phenotype of cancer cells. *Cell* **130,** 986–988 (2007).
229. Frei, E. GENE DELETION - A NEW TARGET FOR CANCER-CHEMOTHERAPY. *Lancet* **342,** 662–664 (1993).
230. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science* **343,** 80–84 (2014).
231. Schreiber, S. L. *et al.* Towards patient-based cancer therapeutics. *Nat. Biotechnol.* **28,** 904–906 (2010).
232. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462,** 108–112 (2009).
233. Cheung, H. W. *et al.* Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 12372–12377 (2011).
234. *RNAi Consortium-shRNA design process*. at <http://www.broadinstitute.org/science/projects/rnai-consortium/trc-shrna-design-process>
235. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391,** 806–811 (1998).
236. Hamilton, A. J. A Species of Small Antisense RNA in Posttranscriptional Gene Silencing in Plants. *Science* **286,** 950–952 (1999).
237. Hammond, S. M., Bernstein, E., Beach, D. & Hannon, G. J. An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature* **404,** 293–296 (2000).
238. Siomi, H. & Siomi, M. C. On the road to reading the RNA-interference code. *Nature* **457,** 396–404 (2009).
239. Yu, J.-Y., DeRuiter, S. L. & Turner, D. L. RNA interference by expression of short-interfering RNAs and hairpin RNAs in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.* **99,** 6047–6052 (2002).
240. Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101,** 25–33 (2000).
241. Schultz, N. *et al.* Off-target effects dominate a large-scale RNAi screen for modulators of the TGF-b pathway andreveal microRNA regulation of TGFBR2. *Silence* **2,** 3 (2011).
242. Reynolds, A. *et al.* Rational siRNA design for RNA interference. *Nat. Biotechnol.* **22,** 326–330 (2004).
243. Gu, S. *et al.* The Loop Position of shRNAsand Pre-miRNAs Is Critical for the Accuracy of Dicer Processing In Vivo. *Cell* **151,** 900–911 (2012).
244. Gu, S. *et al.* Thermodynamic stability of small hairpin RNAs highly influences the loading process of different mammalian Argonautes. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 9208–9213 (2011).
245. Mcintyre, G. J., Yu, Y.-H., Lomas, M. & Fanning, G. C. The effects of stem length and core placement on shRNA activity. *BMC Molecular Biology* **12,** 34 (2011).
246. Jackson, A. L. & Linsley, P. S. Recognizing and avoiding siRNA off- target effects

for target identification and therapeutic application. 1–11 (2010). doi:10.1038/nrd3010
247. Birmingham, A. *et al.* 3′ UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nat Meth* **3,** 199–204 (2006).
248. Sigoillot, F. D. *et al.* A bioinformatics method identifies prominent off-targeted transcripts in RNAi screens. *Nat Meth* **9,** 363–366 (2012).
249. Sigoillot, F. D. & King, R. W. Vigilance and Validation: Keys to Success in RNAi Screening. *ACS Chem. Biol.* **6,** 47–60 (2011).
250. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenicKRAS-driven cancers require TBK1. *Nature* **461,** 108–112 (2009).
251. Cheung, H. W. *et al.* Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 12372–12377 (2011).
252. Luo, B. *et al.* Highly parallel identification of essential genes in cancer cells. *PNAS* 1–6 (2008). doi:10.1073/pnas
253. Shao, D. D. *et al.* ATARiS: computational quantification of gene suppression phenotypes from multisample RNAi screens. *Genome Res.* **23,** 665–678 (2013).
254. Kim, T. M. *et al.* Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Res.* **23,** 217–227 (2013).
255. Li, C. & Wong, W. H. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. U.S.A.* **98,** 31–36 (2001).
256. Li, C. A. W. W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome biology* **2,** research0032.1–research0032.11. (2001).
257. Carter, S. L. M. M. A. G. G. Accurate estimation of homologue-specific DNA concentration ratios in cancer samples allows long-range haplotyping. *Preprint at http://precedings.nature.com/documents/6494/version/1/* (2011).
258. Broad Institute,,. FireHose. *broadinstitute.org* (2011). at <http://www.broadinstitute.org/cancer/cga/Firehose>
259. Stephens, P. J. *et al.* Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell* **144,** 27–40 (2011).
260. Edlundh-Rose, E. *et al.* NRAS and BRAF mutations in melanoma tumours in relation to clinical characteristics: a study based on mutation screening by pyrosequencing. *Melanoma Res.* **16,** 471–478 (2006).
261. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nature Publishing Group* 1–9 (2013). doi:10.1038/ng.2762
262. Fudenberg, G., Getz, G., Meyerson, M. & Mirny, L. A. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat. Biotechnol.* **29,** 1109–U75 (2011).
263. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326,** 289–293 (2009).
264. Artandi, S. E. *et al.* Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. *Nature* **406,** 641–645 (2000).
265. De, S. & Michor, F. nsmb.2089. *Nature Publishing Group* **18,** 950–955 (2011).

266. Nik-Zainal, S. *et al.* The Life History of 21 Breast Cancers. *Cell* **149,** (2012).
267. Yang, L. *et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153,** (2013).
268. Landau, D. A. *et al.* Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell* **152,** 714–726 (2013).
269. Krasnitz, A., Sun, G., Andrews, P. & Wigler, M. Target inference from collections of genomic intervals. *PNAS* (2013). doi:10.1073/pnas.1306909110
270. Mullighan, C. G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446,** 758–764 (2007).
271. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).
272. Solimini, N. L. *et al.* Recurrent Hemizygous Deletions in Cancers May Optimize Proliferative Potential. *Science* **337,** 104–109 (2012).
273. Nijhawan, D. *et al.* Cancer vulnerabilities unveiled by genomic loss. *Cell* **150,** 842–854 (2012).
274. Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26,** i237–i245 (2010).
275. Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
276. Arrowsmith, C. H., Bountra, C., Fish, P. V., Lee, K. & Schapira, M. Epigenetic protein families: a new frontier for drug discovery. *Nature Reviews Drug Discovery* **11,** 384–400 (2012).
277. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. OPTIMIZATION BY SIMULATED ANNEALING. *Science* **220,** 671–680 (1983).
278. Cerny, V. THERMODYNAMICAL APPROACH TO THE TRAVELING SALESMAN PROBLEM - AN EFFICIENT SIMULATION ALGORITHM. *Journal of Optimization Theory and Applications* **45,** 41–51 (1985).
279. Rossin, E. J. *et al.* Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. *PLoS genetics* **7,** (2011).
280. Team, R. C. R: A language and Environment for Statistical Computing. (2012).
281. Jari Oksanen, F. G. B. R. P. L. P. R. M. R. B. O. G. L. S. P. S. M. H. H. S. A. H. W. Vegan: Community Ecology Package. (2012).
282. Raychaudhuri, S. *et al.* Identifying Relationships among Genomic Disease Regions: Predicting Genes at Pathogenic SNP Associations and Rare Deletions. *PLoS genetics* **5,** (2009).
283. Maida, Y. *et al.* An RNA-dependent RNA polymerase formed by TERT and the RMRP RNA. *Nature* **461,** 230–235 (2009).
284. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2015).
285. Du, Z. *et al.* Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nature Structural & Molecular Biology* **20,** 908–– (2013).
286. Cheetham, S. W., Gruhl, F., Mattick, J. S. & Dinger, M. E. Long noncoding RNAs and the genetics of cancer. *British journal of cancer* **108,** 2419–2425 (2013).

287. Ying, H. *et al.* Mig-6 controls EGFR trafficking and suppresses gliomagenesis. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 6912–6917 (2010).
288. Du, J. *et al.* FOXC1, a target of polycomb, inhibits metastasis of breast cancer cells. *Breast Cancer Research and Treatment* **131,** 65–73 (2012).
289. Raychaudhuri, S. *et al.* Identifying Relationships among Genomic Disease Regions: Predicting Genes at Pathogenic SNP Associations and Rare Deletions. *PLoS genetics* **5,** (2009).
290. French, C. A. *et al.* Midline carcinoma of children and young adults with NUT rearrangement. *J. Clin. Oncol.* **22,** 4135–4139 (2004).
291. Kim, T. M. *et al. Genome Res.* **23,** 217–227 (2013).
292. Borrow, J. *et al.* The translocation t(8;l6)(p11, p13) of acute myeloid leukaemia fuses a putative acetyltransferase to the CREB binding protein. *Nat Genet* **14,** 33–41 (1996).
293. Champagne, N. *et al.* Identification of a human histone acetyltransferase related to monocytic leukemia zinc finger protein. *Journal of Biological Chemistry* **274,** 28528–28536 (1999).
294. Jaju, R. J. *et al.* A novel gene, NSD1, is fused to NUP98 in the t(5;11)(q35;p15.5) in de novo childhood acute myeloid leukemia. *Blood* **98,** 1264–1267 (2001).
295. Micci, F., Panagopoulos, I., Bjerkehagen, B. & Heim, S. Consistent rearrangement of chromosomal band 6p21 with generation of fusion genes JAZF1/PHF1 and EPC1/PHF1 in endometrial stromal sarcoma. *Cancer Research* **66,** 107–112 (2006).
296. Park, J. T. *et al.* Notch3 gene amplification in ovarian cancer. *Cancer Research* **66,** 6312–6318 (2006).
297. Garkavtsev, I., Kazarov, A., Gudkov, A. & Riabowol, K. Suppression of the novel growth inhibitor p33(ING1) promotes neoplastic transformation. *Nat Genet* **14,** 415–420 (1996).
298. Beshiri, M. L. *et al.* Coordinated repression of cell cycle genes by KDM5A and E2F4 during differentiation. *PNAS* **109,** 18499–18504 (2012).
299. Gargalionis, A. N., Piperi, C., Adamopoulos, C. & Papavassiliou, A. G. Histone modifications as a pathogenic mechanism of colorectal tumorigenesis. *International Journal of Biochemistry & Cell Biology* **44,** 1276–1289 (2012).
300. Berman, B. P. *et al.* Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat Genet* **44,** 40–U62 (2012).
301. Fullgrabe, J., Kavanagh, E. & Joseph, B. Histone onco-modifications. *Oncogene* **30,** 3391–3403 (2011).
302. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2014).
303. Lowe, S. W. *et al.* P53 STATUS AND THE EFFICACY OF CANCER-THERAPY IN-VIVO. *Science* **266,** 807–810 (1994).
304. Andreassen, P. R., Lohez, O. D., Lacroix, F. B. & Margolis, R. L. Tetraploid state induces p53-dependent arrest of nontransformed mammalian cells in G1. *Mol. Biol. Cell* **12,** 1315–1328 (2001).
305. Ho, C. C., Hau, P. M., Marxer, M. & Poon, R. Y. C. The requirement of p53 for maintaining chromosomal stability during tetraploidization. *Oncotarget* **1,** 583–

595 (2010).
306. Dalton, W. B., Yu, B. & Yang, V. W. p53 suppresses structural chromosome instability after mitotic arrest in human cells. *Oncogene* **29,** 1929–1940 (2010).
307. Tang, Z. Y. *et al.* PP2A is required for centromeric localization of sgol and proper chromosome segregation. *Developmental Cell* **10,** 575–585 (2006).
308. Khanna, K. K. & Jackson, S. P. DNA double-strand breaks: signaling, repair and the cancer connection. *Nat Genet* **27,** 247–254 (2001).
309. Filippova, G. N. *et al.* Tumor-associated zinc finger mutations in the CTCF transcription factor selectively alter its DNA-binding specificity. *Cancer Research* **62,** 48–52 (2002).
310. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** (2012).
311. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology* **12,** R41 (2011).
312. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607 (2012).
313. Zhang, R. & Lin, Y. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res.* **37,** D455–D458 (2009).
314. Bell, D. *et al.* Integrated genomic analyses of ovarian carcinoma. *Nature* **474,** 609–615 (2011).
315. Zhang, R. & Lin, Y. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res.* **37,** D455–8 (2009).
316. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607 (2012).
317. Bell, D. *et al.* Integrated genomic analyses of ovarian carcinoma. *Nature* **474,** 609–615 (2011).
318. Ortiz-Estevez, M., las Rivas, De, J., Fontanillo, C. & Rubio, A. Segmentation of genomic and transcriptomic microarrays data reveals major correlation between DNA copy number aberrations and gene-loci expression. *Genomics* **97,** 86–93 (2011).
319. Fu, X. *et al.* Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* **10,** 161 (2009).
320. Marcotte, R. *et al.* Essential Gene Profiles in Breast, Pancreatic, and Ovarian Cancer Cells. *Cancer Discovery* **2,** 172–189 (2012).
321. Schoenheiner, R. *The dynamic state of body constituents*. (Harvard University Press, 1942).
322. SIMPSON, M. V. The release of labeled amino acids from the proteins of rat liver slices. *Journal of Biological Chemistry* **201,** 143–154 (1953).
323. Goldberg, A. L. Degradation of abnormal proteins in Escherichia coli (protein breakdown-protein structure-mistranslation-amino acid analogs-puromycin). *Proc. Natl. Acad. Sci. U.S.A.* **69,** 422–426 (1972).
324. Goldberg, A. L. & Dice, J. F. Intracellular protein degradation in mammalian and bacterial cells. *Annu. Rev. Biochem.* **43,** 835–869 (1974).
325. Hershko, A., Ciechanover, A. & Rose, I. A. Resolution of the ATP-dependent proteolytic system from reticulocytes: a component that interacts with ATP.

Proc. Natl. Acad. Sci. U.S.A. **76,** 3107–3110 (1979).
326. Etlinger, J. D. & Goldberg, A. L. A soluble ATP-dependent proteolytic system responsible for the degradation of abnormal proteins in reticulocytes. *Proc. Natl. Acad. Sci. U.S.A.* **74,** 54–58 (1977).
327. Hershko, A., Heller, H., Elias, S. & Ciechanover, A. Components of ubiquitin-protein ligase system. Resolution, affinity purification, and role in protein breakdown. *Journal of Biological Chemistry* **258,** 8206–8214 (1983).
328. Hershko, A., Ciechanover, A., Heller, H., Haas, A. L. & Rose, I. A. Proposed role of ATP in protein breakdown: conjugation of protein with multiple chains of the polypeptide of ATP-dependent proteolysis. *Proc. Natl. Acad. Sci. U.S.A.* **77,** 1783–1786 (1980).
329. Tanaka, K., Mizushima, T. & Saeki, Y. The proteasome: molecular machinery and pathophysiological roles. *Biological Chemistry* **393,**
330. Kaneko, T. *et al.* Assembly Pathway of the Mammalian Proteasome Base Subcomplex Is Mediated by Multiple Specific Chaperones. *Cell* **137,** 914–925 (2009).
331. Thompson, D., Hakala, K. & DeMartino, G. N. Subcomplexes of PA700, the 19 S regulator of the 26 S proteasome, reveal relative roles of AAA subunits in 26 S proteasome assembly and activation and ATPase activity. *The Journal of biological chemistry* **284,** 24891–24903 (2009).
332. Roelofs, J. *et al.* Chaperone-mediated pathway of proteasome regulatory particle assembly. *Nature* **459,** 861–865 (2009).
333. Saeki, Y., Toh-e, A., Kudo, T., Kawamura, H. & Tanaka, K. Multiple Proteasome-Interacting Proteins Assist the Assembly of the Yeast 19S Regulatory Particle. *Cell* **137,** 900–913 (2009).
334. Glickman, M. H., Rubin, D. M., Fried, V. A. & Finley, D. The Regulatory Particle of the Sacharomyces cerevisiae proteasome. *Molecular and Cellular Biology* **18,** 3149–3162 (1998).
335. Meiners, S. Inhibition of Proteasome Activity Induces Concerted Expression of Proteasome Genes and de Novo Formation of Mammalian Proteasomes. *The Journal of biological chemistry* **278,** 21517–21525 (2003).
336. Xie, Y. Feedback regulation of proteasome gene expression and its implications in cancer therapy. *Cancer Metastasis Rev* **29,** 687–693 (2010).
337. Xie, Y. & Varshavski, A. RPN4 is a ligand, substrate, and transcriptionalregulator of the 26S proteasome: A negativefeedback circuit. *PNAS* **98,** 3056–3061 (2001).
338. Goldberg, A. L. Protein degradation and protection against misfolded or damaged proteins. *Nature* **426,** 895–899 (2003).
339. Obeng, E. A. Proteasome inhibitors induce a terminal unfolded protein response in multiple myeloma cells. *Blood* **107,** 4907–4916 (2006).
340. Wu, W. K. K. *et al.* Cancer Letters. *Cancer Letters* **293,** 15–22 (2010).
341. Lee, A.-H., Iwakoshi, N. N., Anderson, K. C. & Glimcher, L. H. Proteasome inhibitors disrupt the unfolded protein response in myeloma cells. *Proc. Natl. Acad. Sci. U.S.A.* **100,** 9946–9951 (2003).
342. Orlowski, R. Z. Phase I Trial of the Proteasome Inhibitor PS-341 in Patients With Refractory Hematologic Malignancies. *J. Clin. Oncol.* **20,** 4420–4427

(2002).
343. Richardson, P. G. *et al.* Bortezomib or high-dose dexamethasone for relapsed multiple myeloma. *N. Engl. J. Med.* **352,** 2487–2498 (2005).
344. Liu, J. *et al.* A genetically defined model for human ovarian cancer. *Cancer Research* **64,** 1655–1663 (2004).
345. Ren, Y. *et al.* Targeted tumor-penetrating siRNA nanocomplexes for credentialing the ovarian cancer target ID4. *Science Translational Medicine* **In press,** (2012).
346. Weinstein, I. B. & Joe, A. K. Mechanisms of disease: Oncogene addiction--a rationale for molecular targeting in cancer therapy. *Nature clinical practice. Oncology* **3,** 448–457 (2006).
347. Ling, Y. H. *et al.* PS-341, a novel proteasome inhibitor, induces Bcl-2 phosphorylation and cleavage in association with G(2)-M phase arrest and apoptosis. *Mol. Cancer Ther.* **1,** 841–849 (2002).
348. Yin, D. *et al.* Proteasome inhibitor PS-341 causes cell growth arrest and apoptosis in human glioblastoma multiforme (GBM). *Oncogene* **24,** 344–354 (2005).
349. Elsasser, S., Schmidt, M. & Finley, D. Characterization of the proteasome using native gel electrophoresis. *Methods in enzymology* **398,** 353–363 (2005).
350. Kisselev, A. F. & Goldberg, A. L. Monitoring activity and inhibition of 26S proteasomes with fluorogenic peptide substrates. *Methods in enzymology* **398,** 364–378 (2005).
351. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483,** 570–575 (2012).
352. Funakoshi, M., Tomko, R. J., Kobayashi, H. & Hochstrasser, M. Multiple Assembly Chaperones Govern Biogenesis of the Proteasome Regulatory Particle Base. *Cell* **137,** 887–899 (2009).
353. Park, S. *et al.* Hexameric assembly of the proteasomal ATPases is templated through their C termini. *Nature* **459,** 866–U9 (2009).
354. Moellering, R. E. *et al.* Direct inhibition of the NOTCH transcription factor complex. *Nature* **462,** 182–188 (2009).
355. Darnell, J. E. Transcription factors as targets for cancer therapy. *Nat Rev Cancer* **2,** 740–749 (2002).
356. WANG, H., HAN, H., MOUSSES, S. & VONHOFF, D. Targeting Loss-of-Function Mutations in Tumor-Suppressor Genes as a Strategy for Development of Cancer Therapeutic Agents. *Seminars in Oncology* **33,** 513–520 (2006).
357. Katayama, R. *et al.* Mechanisms of Acquired Crizotinib Resistance in ALK-Rearranged Lung Cancers. *Science Translational Medicine* **4,** 120ra17–120ra17 (2012).
358. Lu, S. *et al.* Point Mutation of the Proteasome 5 Subunit Gene Is an Important Mechanism of Bortezomib Resistance in Bortezomib-Selected Variants of Jurkat T Cell Lymphoblastic Lymphoma/Leukemia Line. *Journal of Pharmacology and Experimental Therapeutics* **326,** 423–431 (2008).
359. Flaherty, K. T. *et al.* Inhibition of mutated, activated BRAF in metastatic melanoma. *N. Engl. J. Med.* **363,** 809–819 (2010).
360. Kaneko, T. *et al.* Assembly Pathway of the Mammalian Proteasome Base

Subcomplex Is Mediated by Multiple Specific Chaperones. *Cell* **137,** 914–925 (2009).
361. Aghajanian, C. *et al.* A phase I trial of the novel proteasome inhibitor PS341 in advanced solid tumor malignancies. *Clinical cancer research : an official journal of the American Association for Cancer Research* **8,** 2505–2511 (2002).
362. Davoli, T. *et al.* Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome. *Cell* 1–15 (2013). doi:10.1016/j.cell.2013.10.011
363. Carter, S. L., Eklund, A. C., Kohane, I. S., Harris, L. N. & Szallasi, Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat Genet* **38,** 1043–1048 (2006).
364. Fong, P. C. *et al.* Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. *N. Engl. J. Med.* **361,** 123–134 (2009).
365. Muller, F. L. *et al.* Passenger deletions generate therapeutic vulnerabilities in cancer. *Nature* **488,** 337–342 (2013).
366. Wahl, M. C., Will, C. L. & LUhrmann, R. The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell* **136,** 701–718 (2009).
367. Barrangou, R. *et al.* CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* **315,** 1709–1712 (2007).
368. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).
369. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
370. MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42,** D986–D992 (2013).
371. Monti, S. *et al.* DNA copy number inference pipeline forAffymetrix SNP6.0 arrays. *Written but unpublished* 1–8 (2009).
372. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23,** 657–663 (2007).